Single-Photon Techniques for Standard CMOS Digital ICs

THÈSE N^{O} 4954 (2011)

PRÉSENTÉE LE 13 AVRIL 2011

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

GROUPE CHARBON

PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Claudio FAVI

acceptée sur proposition du jury:

Prof. G. De Micheli, président du jury
Prof. E. Charbon, directeur de thèse
Dr P. Jarron, rapporteur
Dr M. Mattavelli, rapporteur
Dr L. Wang, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2011

CONTENTS

RÉSUMÉ

L'arrivée des photodétecteurs à photons individuels connus sous l'acronyme de SPAD (Single-Photon Avalanche Diode) dans un procédé de fabrication CMOS standard, a ouvert de nouvelles perspectives dans l'intégration de ces capteurs ultra-sensibles avec de la logique digitale.

La lumière a des propriétés intéressantes qui attirent les chercheurs en informatique et électronique depuis longtemps. Sa nature ondulatoire et donc sans masse fait d'elle une candidate idéale pour remplacer les électrons lorsque cela n'est pas déjà le cas. Ceci est notamment le cas pour les transferts de données à longue distance. Cependant une nouvelle tendance peut être observée. Les chercheurs se tournent vers la photonique, la science de la lumière, pour communiquer également à courte distance. Une des raisons de cette tendance est l'efficacité accrue en consommation énergétique. En effet, les photons, les particules élémentaires de la lumière, à l'instar des électrons, ne subissent pas d'effets résistifs, capacitifs ou inductifs. La nature ondulatoire de la lumière permet aussi aux concepteurs de se libérer de problèmes tels que le *cross-talk* et l'injection de bruit tout en tirant meilleur parti des phénomènes d'interférence. Cependant, la photonique n'est pas encore la panacée et nous ne pensons pas qu'elle remplacera entièrement un jour l'électronique traditionnelle. Une combinaison des deux sera très certainement adoptée afin de profiter au mieux des deux mondes.

La conception de circuits intégrés digitaux en technologie CMOS fait face à de nombreux obstacles et il n'est pas clair si la technologie sera encore capable, dans le futur, de fournir des réductions de taille, des performances accrues et des consommations réduites. Dans ce travail, nous présentons trois investigations où la photoniques à base de SPADs est intégrée avec la technologie digitale CMOS.

Les trois contributions principales de cette thèse sont: des paradigmes de communication à base de photons individuels, des techniques de traitement et de lecture de capteurs à photons individuels, et des méthodes de distribution d'horloge et de synchronisation basées sur les SPADs. La communication à base de photons individuels est proposée en combinant des SPADs avec des

TDCs ultra-rapides avec une modulation dédiée. Dans ce contexte, les limites théoriques de la capacité du canal en présence de bruit et d'autres sources de non-uniformité liées aux SPADs ont été dérivées; un TDC d'une résolution de 17 ps a été démontré sur une plateforme FPGA. A ce jour et au mieux de nos connaissances, ceci est la plus haute résolution reportée pour ce type de TDC. Le traitement et la lecture de capteurs à photons individuels ont été démontrés dans plusieurs technologies, et en particulier avec des systèmes d'imagerie à photons individuels où des architectures massivement parallèles ont été étudiées et implémentées en CMOS. Des méthodes de distribution d'horloge et de synchronisation pouvant potentiellement éliminer le problème de *skew* indépendamment de la taille d'un chip ont été démontrées. Les avantages en termes d'efficacité de cette approche sont particulièrement intéressants dans les systèmes embarqués avec extension du jeu d'instructions.

Cette thèse emploie la technologie des SPADs en CMOS pour plusieurs applications en créant ainsi un pont entre la conception de systèmes digitaux et la photonique à haute performance. A notre connaissance, ceci est la première tentative dans cette direction visant la technologie CMOS.

Mots clés: Single-Photon Avalanche Diode (SPAD), digital CMOS, single-photon communication, photon channel capacity, Time-to-Digital Converter (TDC), single-photon imaging, photon processing, Time-Uncorrelated Photon Counting (TUPC), Time-Correlated Single-Photon Counting (TCSPC), sensor array readout strategies, single-photon clocking, clock distribution, clock networks.

ABSTRACT

The advent of single-photon detectors known as Single-Photon Avalanche Diodes in standard CMOS technology opened the way to new perspectives in integrating these ultra sensitive light sensors with digital logic.

Light has some interesting properties that attracted researchers in computer and electronics for a long time. Its weightlessness nature makes it a candidate to replace electrons when it didn't already do so. This is particularly true for long distance data transfers. However a new trend can be observed. Researchers are looking into photonics, the science of light, for short distance communications as well. Power efficiency is one of the reasons of this trend. In fact, photons, the elementary particles of light, don't suffer of resistive, capacitive, or inductive effects like electrons do. The wavelike nature of light can also free designers from problems such as cross-talk and side-channel noise injection while taking advantage of interference. However photonics is still not the panacea and we don't believe it will ever completely replace electronics. Most certainly a combination of the two will be adopted to take advantage of both worlds.

CMOS digital Integrated Circuit (IC) design faces many challenges and it is not clear whether technology will still provide small footprints, high performance, and low power in the future. Photonics with SPADs can answer some of these questions. In this work, we present three investigations where SPAD photonics is integrated within digital CMOS technology.

The main contributions of this thesis are threefold: single-photon CMOS communication paradigms, single-photon processing and readout techniques, single-photon clocking and synchronization methods. Single-photon communication was achieved using a combination of SPADs and ultra-fast TDCs in a pulse position modulation scheme. In this context, theoretical channel capacity limits in the presence of noise and other non-idealities typical of SPADs were derived; a TDC with a resolution of 17 ps was demonstrated in a standard FPGA fabric. To the best of our knowledge, at the time of this writing, this is the highest reported resolution for a TDC of this kind. Single-photon processing and readout was achieved in several technologies, focusing on im-

age sensor design, whereby massive parallel architectures were studied and implemented in CMOS. Single-photon clocking and synchronization was demonstrated allowing potentially zero skew systems irrespective of the chip area. The power benefits of this approach in embedded systems with instruction set extensions are particularly interesting.

The thesis makes use of SPAD technology implemented in CMOS for a number of applications creating a bridge between digital design and high performance photonics. We believe that this is the first attempt in this direction focused on CMOS technology.

Keywords: Single-Photon Avalanche Diode (SPAD), digital CMOS, single-photon communication, photon channel capacity, Time-to-Digital Converter (TDC), single-photon imaging, photon processing, Time-Uncorrelated Photon Counting (TUPC), Time-Correlated Single-Photon Counting (TCSPC), sensor array readout strategies, single-photon clocking, clock distribution, clock networks.

1 INTRODUCTION

COMPUTING has become the most pervasive resource of our daily lives. Computing devices have become more and more powerful while size and electrical consumption have shrunk. The most shocking comparison is perhaps that of our mobile phones that pack, in a few squared centimeters, more processing power than supercomputers only a few decades ago, at a fraction of the power. The chips that power our devices are the fruit of cunningly tailored trade-offs between performance, power, and die-size (i.e. cost). Very Large Scale Integration (VLSI) was a widely employed keyword a few years ago. Today, almost every electronic device produced can be labeled as a highly integrated technology.

High integration is a key factor in mobile applications such as mobile phones, laptops, and more recently, tablet computers. Small footprint devices also spawned from the need of higher inter-chip bandwidth, which is generally limited by capacitive and inductive effects. Specialized devices, such as MEMs, flash memories, or optical stacks, that require specific fabrication technology have also become integrated in systems known as Systems-on-Chip (SoC). Scaling in size of the electrical interface of integrated circuits, also known as pads, has not followed the transistors' trend. In fact, wire-bonding and flip chip techniques are limiting the pads' size reduction.

Low power operation is not restricted to mobile devices; power budgets and, more specifically, thermal extraction limits are real constraints even in high-performance applications. For example, data centers packed with racks of commodity or high-end computing platforms have enormous power requirements for normal operation and air conditioning. While the cost of operation can be

reduced for the customer, the eco-friendly arguments is also becoming a non-negligible incentive.

High performance, at the risk of stating the obvious, is pursued to enable evermore computing/networking intensive applications. Performance can be characterized in terms of throughput, bandwidth, latency, and availability, to name a few usual metrics. When raw performance cannot be achieved by mere scaling of technology and clock frequency, parallelism is employed at the cost of silicon area. In fact, the area of chips keeps growing as more cores are integrated and heterogeneous systems created.

These three aforementioned contradictory requirements (high integration, low power, and high performance) are inevitably present in today's devices design. Technology scaling has helped pushing these limits, nonetheless, researchers are investigating alternative methods in order to expand beyond the current limitations.

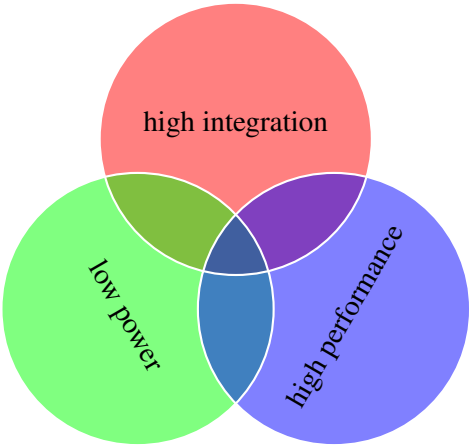


Figure 1.1: Tradeoff in current chip design

.

One alternative method is to use photons, the elementary particle and fundamental unit of light. Photonics has already been exploited in the past for long distance communication. For a long time and still to date sensors have been made in custom processes that are generally not compatible with the standard logic process widely used and known as Complementary Metal-Oxide Semiconductor (CMOS) technology. In 2003, the horizon changed dramatically when Rochas *et al.* presented the first single-photon detector fabricated in CMOS technology [1]. In fact, starting around the beginning of the 21st century, a major focus of photonics has been on silicon. Silicon being the element used as substrate in CMOS technology, the integration of single-photon sensors in CMOS opened new applications where light sensing and digital data processing are combined.

Silicon photonics research at Intel [2] and IBM [3] for example, are now focusing on building light emitting devices, modulating devices, and receivers for chip-to-chip communication in the scope of high performance computing at first. In board-to-board and cabinet level interconnect, optical communication is the *de facto* standard. These research programs are pushing the boundary to the chip level and they are also addressing board level optical transmission [4].

The scope of this thesis is to propose an optical approach where the electrical one falls short and, by no means, promote an all-optical solution. Also the single-photon techniques presented do not necessarily imply that only a single photon is used but rather that a single detected photon is sufficient to trigger the desired effect. In fact, for practical reasons but to some extent, many photons can be used.

After a review of the existing state of the art in CMOS based photon detection, generation, and transport in chapter 2, the thesis is organized in three main sections. First, single-photon communication amenable for inter- and intra-chip communication is discussed

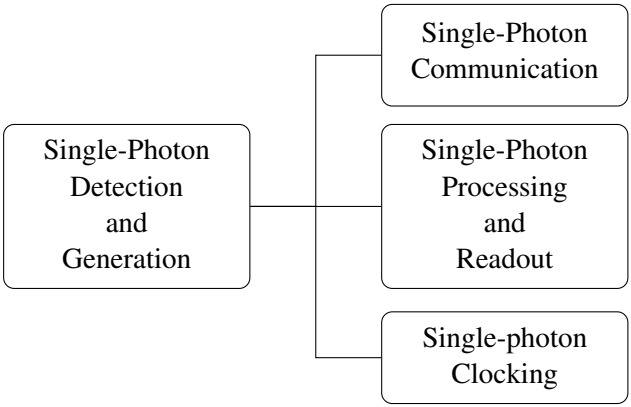


Figure 1.2: Structure of the thesis.

in chapter 3. A time-to-digital converter implementation and evaluation is particularly detailed in this chapter. Second, chapter 4 discusses single-photon processing, focusing on single-photon detection and readout integration in CMOS, as well as its applications. Third, a system for optical clock distribution is presented in chapter 5. The concluding chapter 6 describes future work and other possible applications. The main contributions of this work are also summarized in chapter 6.

2 Single-Photon Detection and Generation

PHOTONICS is the science of light. From generation to detection, it covers amongst others transmission, modulation, and amplification of light. Photons, the elementary particles of light, present the particular duality of being both particles and waves. Although we will focus on the corpuscular aspect mainly, the wave-like aspects should be kept in mind.

Silicon photonics is a sub-class of photonics in which the light interacts with the most widely used semiconductor in industry: silicon. The semiconductor industry has grown and refined processes to fabricate larger and larger chips on purified crystalline silicon dies with very well controlled yield. High integration again is the motive that drives research of photonics with silicon. The major players, Intel and IBM, have silicon photonics research programmes [2, 3] that have produced very interesting results these recent years.

This chapter is meant as a review of the state of the art in silicon photonics with a certain bias on detectors. However we will shortly review light generation devices in section 2.2, and photons transport and modulation techniques in silicon in section 2.3.

2.1 Single-Photon Detection

Originally, detection of single-photon events has been and is still performed with Photomultiplier Tubes (PMTs). These vacuum tube devices, generally bulky by construction, provide detection in a wide spectrum with low noise but with high time response and jitter in the order of nanoseconds. Multichannel or Microchannel

Plates (MCPs), while still relatively large, can also provide opto-electric conversion and amplification with multiple channels that can be used for example in event position detection. Multichannel Plates (MCPs) may reach picosecond timing resolutions.

Avalanche Photodiodes (APDs) are solid state devices in which photon-induced electron-hole pairs are accelerated by an electric field. These accelerated carriers may produce secondary electron-hole pairs by impact ionization. This process is known as avalanche multiplication. APDs have been built in GaAs, SiGe, and Si substrates. The first appearing APDs, also known as reach-through APDs (RAPDs), are thick devices built "vertically" in which a large region of absorption (π region) is layered on top of a p-n multiplication region. RAPDs have high breakdown voltage, high sensitivity in the visible and IR spectrum, large active area, and relatively poor timing resolution. A second category of APDs, which are the subject of interest of this thesis, are thin devices that present a reduced active area, low photon detection but excellent timing resolution are built with planar technology. Moreover, another advantage of the planar technology is that it suits well to high densitiy integration and quenching circuitry when required for normal operation.

A Single-Photon Avalanche Diode (SPAD) is an APD operated above breakdown voltage, in the so-called Geiger mode. In Geiger mode of operation, SPADs exhibit a virtually infinite optical gain, however a mechanism must be provided to quench the avalanche.

SPADs and APDs in general have been built in custom and rather expensive processes until Rochas *et al.* presented the first SPAD built in standard CMOS technology, in 2003 [1, 5]. The breakthrough opened the way to large integrated arrays of SPADs [6] with an extremely low fabrication cost due to the use of a standard process. Many efforts have been made to port SPADs from sub-micron technology to deep sub-micron technology, such as [7] for 0.35 µm, [8–10] for 180 nm, [11–13] for 130 nm, and [14] for

90 nm technology. Generally speaking SPADs show excellent timing response and good quantum efficiency at the expense of long dead time and hard-to-control noise rate figures.

2.1.1 Single-Photon Avalanche Diode

The SPAD active region or depletion region is a p-n junction that forms the diode. Biasing the diode near or above breakdown voltage induces a high electric field in the depletion region enabling the avalanche multiplication process. A typical I-V curve of a SPAD is shown in figure 2.1, the breakdown voltage for this device is 9.4 V.

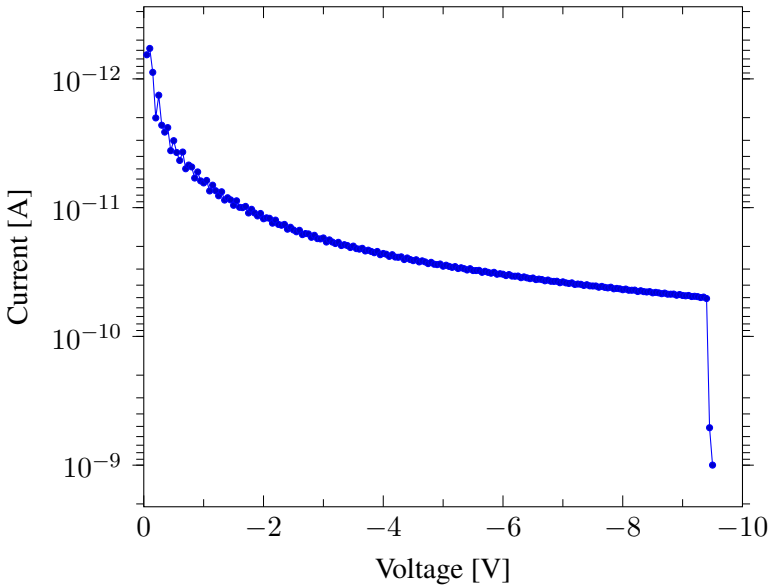


Figure 2.1: I-V curve of a 130 nm SPAD [12, 13].

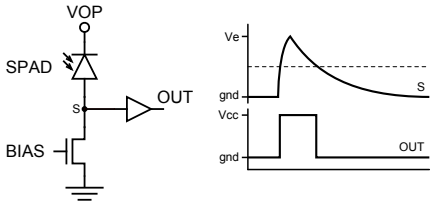


Figure 2.2: The SPAD and surrounding buffering and quenching circuitry.

The operating voltage V_{op} is generally described in terms of the breakdown voltage V_{bd} of the device and the excess bias V_{e}:

$$V_{\mathrm{op}} = |V_{\mathrm{bd}}| + V_{\mathrm{e}}.$$

The device requires quenching circuitry in order to stop the avalanche-induced current. Failure to do so would damage the device. There exist several techniques to accomplish quenching, classified in active and passive quenching. The simplest approach to passive quenching is to use a ballast resistance. The avalanche current causes the diode reverse bias voltage to drop below breakdown, thus pushing the junction to linear avalanching and even pure accumulation mode. Active quenching generally consists of a feedback-forced voltage drop below breakdown with the same effects as passive quenching. After quenching, the device requires a certain recovery or recharge time, to return to the initial state. Recharge can also be active or passive depending whether the bias of the diode is forced or not forced above breakdown. The quenching and recovery times are collectively known as dead time. Figure 2.2 shows a passive quenching and passive recharge scheme.

Hereafter will be discussed several figures of merit of SPADs such as DCR, PDP, jitter, dead time, and afterpulsing. A comparison of several devices will be also given toward the end of this section. Knowing the fundamental limitations of these devices yields

to several applications that can take advantage of the technology while avoiding their drawbacks.

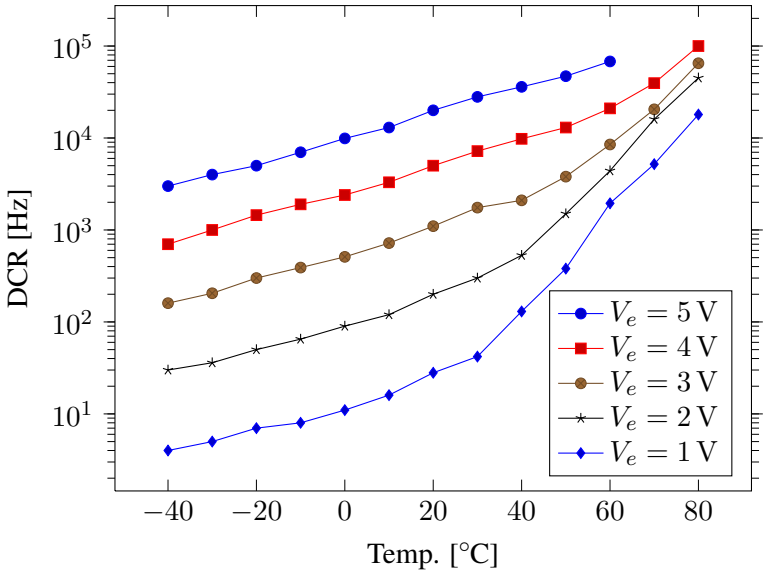


Figure 2.3: Example of dark count rate versus temperature with a 130 nm SPAD. Source: [13].

SPADs noise measured as Dark Count Rate (DCR) is the mean frequency of spurious pulses. Spurious pulses are mainly due to tunneling and thermal generation. Tunneling in SPADs is a stochastic phenomenon where a particle crosses the bandgap. The tunneling probability highly increases with the electric field and becomes dominant for field values in excess of 10^6 V/cm. Thermal generation can cause an electron in the valence band to transition to the conduction band. The presence of *traps*, intermediate energy levels in the bandgap, accentuate both tunneling and thermal generation. Figure 2.3 shows an example of DCR variation as a function of tem-

perature for a 130 nm CMOS implementation. Also shown in the figure, the DCR also increases as V_e increases. Note that different devices may show different DCR behaviors depending on whether tunneling or thermal effects are dominant. For the case depicted in figure 2.3, we can see that for temperatures above 40 °C thermally generated free carriers are the main contributors, while tunnelling dominates at room temperature and below.

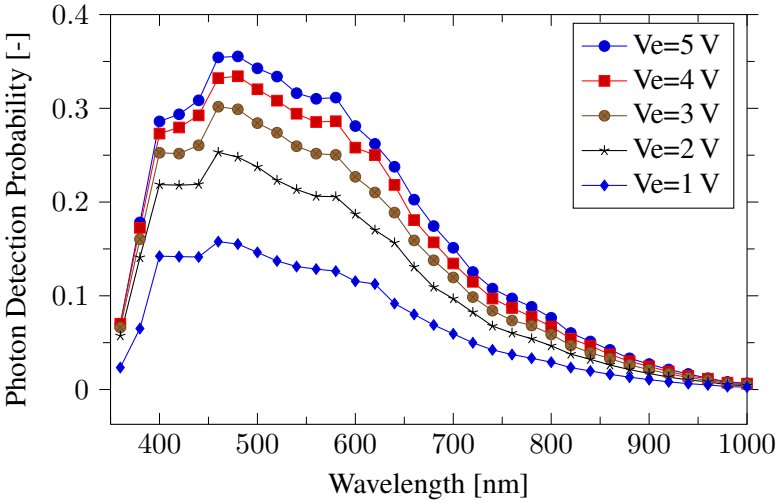


Figure 2.4: Illustration of Photon Detection Probability curves at different excess bias. Source: [12, 13].

Photon Detection Probability (PDP) is defined as the probability that a photon of wavelength λ generates a pulse at the output of the SPAD. For CMOS SPADs there are two factors that affect PDP. First, the photons need to reach the silicon and be absorbed. Depending the optical stack, photons are reflected, refracted, or absorbed before reaching the substrate. Second, photons that reach

the substrate must generate an electron-hole pair for an avalanche build-up to be triggered. For this to happen, the generated free carriers must lie in the depletion region of the SPAD (as carriers generated in a region with a low electric field will very likely recombine). Therefore the depth of the depletion region and the excess bias play an important role in PDP figures. Figure 2.4 shows PDP curves for a 130 nm CMOS SPAD. Note that increasing excess bias V_e does not always increase PDP as at some point DCR will dominate hence PDP decreases. Note that CMOS SPADs have a peak in PDP between 400 nm and 600 nm.

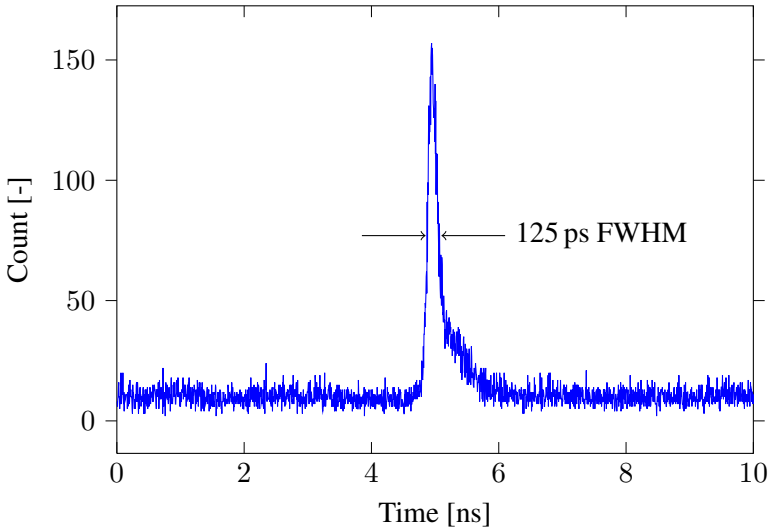


Figure 2.5: Histogram of the timing of pulses generated by SPAD illuminated with a fixed frequency picosecond laser source [12, 13].

Timing resolution or *jitter* is crucial for application where the exact arrival time of photons (and the event they represent) is required. The timing resolution in SPADs is influenced by the avalan-

che build-up process, parasitic capacitance, and the noise threshold of the output buffer. The avalanche steepness is dependent on the electrical field, while the capacitance and the threshold noise are dependent of the geometry and the technology. SPADs jitter is characterized by illuminating the active region with a fixed-frequency low-jitter laser source and controlling the intensity so as to prevent pileup effects. An histogram of the arrival time of SPAD's pulses with respect to the laser trigger is constructed. The full-width-at-half-maximum value is taken as timing jitter. Figure 2.5 shows an example of histogram built with a 40 MHz laser diode. Note that the measurement taken also accounts for jitter in the laser source and the measurement instrument both of which are relatively small (a few picoseconds) compared to the measured jitter.

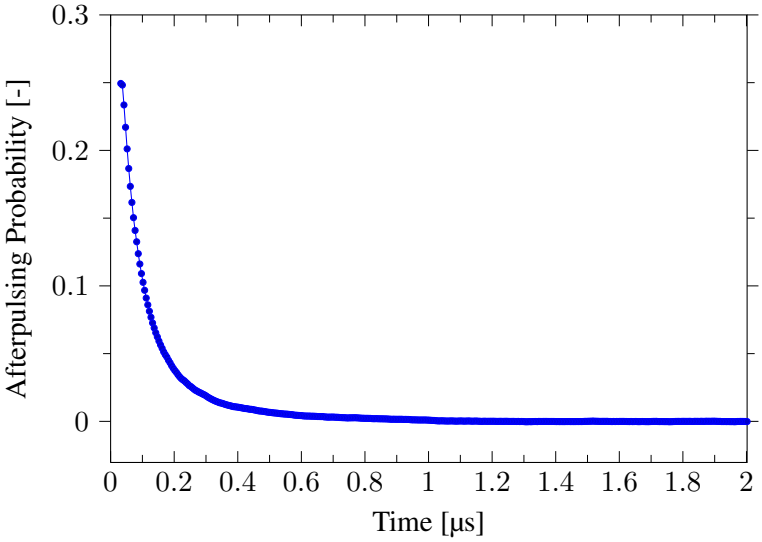


Figure 2.6: Afterpulsing probability at nominal dead-time (40 ns) for a 0.35 µm SPAD.

The *dead time* is defined as the minimum time between two photons that the sensor can unambiguously detect. For SPADs, the dead time is highly dependent on the quenching and, if present, recharging circuitry. The dead time is a design parameter, when building a SPAD with the surrounding circuitry (also known as SPAD ensemble), that affects the Afterpulsing Probability (APP). Afterpulses are avalanche breakdowns that occur when carriers from previous avalanches get trapped in the multiplication region. Afterpulsing effects can be reduced by either limiting the number of carriers an avalanche builds or allowing trapped carriers to evacuate the multiplication region before the SPAD is recharged to the operational state. Carrier limitation is done by reducing the parasitic capacitances. In CMOS, the integrated quenching and read-out circuitry intrinsically limit parasitic capacitances. Slowing down the recharge time, hence increasing the dead time, contributes to lower the APP. Figure 2.6 shows a typical APP curve.

Table 2.1 presents a comparison of CMOS SPADs from the first 0.8 µm device presented by Rochas *et al.* [1] to the latest 90 nm SPAD by Karami [14, 15]. The trend of the decreasing breakdown voltage (from the older to the newer devices) is due to increasing doping levels and increased tunneling effects [16]. The DCR increase is also mainly due to tunneling. The large timing jitter of [15] can be attributed to the large active area combined with the diffusion of minority carriers, generated by photons absorbed deeper than the depletion region, back into the multiplication region. The depletion region being shallower in this technology due to a higher electric field.

Performance measure	[15]	[13]	[11]	[10]	[9]	[8]	[7]	[6]	[1]	Unit
Technology	90	130	130	180	180	180	350	800	800	nm
Max. PDP	12	18 [a]	34 [a]	N.A.	2.5[a]	14 [a]	40 [a]	26 [a]	28 [a]	%
Max. PDP	N.A.	30 [b]	41 [b]	N.A.	5.5[b]	N.A.	35 [b]	N.A.	33 [b]	%
Typ. DCR	16	90 [a]	95 [a]	N.A.	70 [a]	0.1[a]	N.A.	0.35[a]	0.9[a]	kHz
Typ. DCR	N.A.	670 [b]	950 [b]	∼1000 [b]	N.A.	N.A.	0.75[b]	N.A.	4 [b]	kHz
Active area	50	58	78	3.1	∼11	78	10	38	38	μm^2
Timing jitter (FWHM)	398	125	144	N.A.	N.A.	N.A.	80	115	60	ps
Afterpuling probability	32	<1	N.A.	N.A.	N.A.	N.A.	23	N.A.	8	%
Dead time	N.A.	180	N.A.	N.A.	30	N.A.	40	40	75	ns
Breakdown voltage	10.4	9.4	9.9	11	10.2	10	17.7	25.5	25.25	V
[a]Excess Bias (Ve)	0.15	1	1.7	N.A	0.5	0.5	3.3	5	5	V
[b]Excess Bias (Ve)	N.A.	2	2.7	2.5	2	N.A.	4	N.A.	10	V

Table 2.1: Performance comparison of SPAD implementations.

2.2 Photon Generation

Light generation can be categorized in many ways: incadescence for light emission of hot body, chemiluminescence for light emission in chemical reactions, electroluminescence for light emission in response to an electric current, mechanoluminescence for light emission by solid under mechanical stress, etc. We are interested in electroluminescence as it can be conveniently controlled electrically, however the principles of light emission are similar in many other cases. Two fundamental photon emission cases are known: spontaneous emission and stimulated emission.

In the first case, an electron in an excited state at energy level E_h may transition to a lower level E_l. While doing so, it emits a photon with energy $h\nu = E_h - E_l$, where ν is the frequency and h Planck's constant (fig. 2.7a). Spontaneous emission is fundamental process behind the incandescent light bulbs, light-emitting diodes, cathodic or plasma displays to cite a few.

In stimulated emission, photons with energy $h\nu$ are generated by the transition of an electron from an excited level to a ground level similar to spontaneous emission. However, the transition is triggered by another photon of energy $h\nu$ (fig. 2.7b). Note also that the newly generated photon's phase, frequency, polarization, and direction are the same as the triggering photon. The process of stimulated emission is used principally in lasers.

The integration of spontaneous and stimulated emission devices is a highly researched area. Because of the indirect bandgap of silicon, highly integrated spontaneous emission devices often use III-V nitride direct bandgap semiconductors processed on top sapphire substrates [17–21] that can be bump-bonded to CMOS drivers for instance [22]. Efficient silicon LEDs were demonstrated by using dislocation loops [23] or one- and two-photon assisted sub-bandgap light emission [24]. Experiments in fluorescence life-time imag-

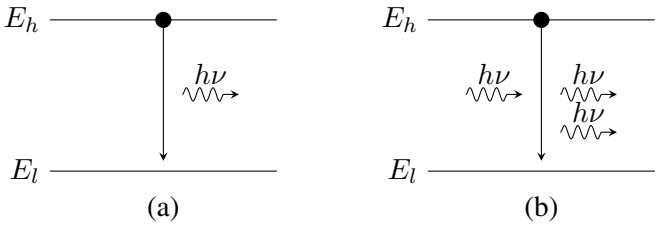


Figure 2.7: Spontaneous photon emission (a) when an electron a excited state E_h transits to a lower lever energy state. In stimulated emission (b) the transition is triggered by a existing photon, the newly emitted photon has the same phase, frequency, polarization, and direction as the the triggering photon.

ing (FLIM) are good examples of use of integrated micro-LEDs arrays [22] while optical stack considerations are shown in [20]. SPADs also present spontaneous emission of photons during avalanche breakdown although the intensity is rather limited and the timing not yet well controlled [13].

Lasers are stimulated emission devices that are known since the 1960s. Semiconductor lasers are today the most produced laser found in many consumer electronic apparatuses. Similarly to LEDs, silicon integrated lasers were only first demonstrated in 2004 [25] and are still an active subject of research [26, 27], the main limitation of these devices being their large size and require external optical pumping. Hybrid lasers such as [28] address theses issues by incorporating III-V elements in the amplification region and coupling light in an SOI cavity.

2.3 Photon Transport and Modulation

Free air transport of photons can be the first approach to moving light from point A to point B. However, scattering, reflections, and refractions might render an optical setup difficult to control. Confinement of light can be done efficiently with optical waveguides. By taking advantage of total internal reflection between two material of different refractive index, light is confined and generally can be channeled in a given direction with low attenuation. Optical fibers generally made of silica (SiO_2) use this property for long distance communication.

Low attenuation is dependent on the wavelength and the material used. For example, the absorption coefficients of silicon are shown in figure 2.8.

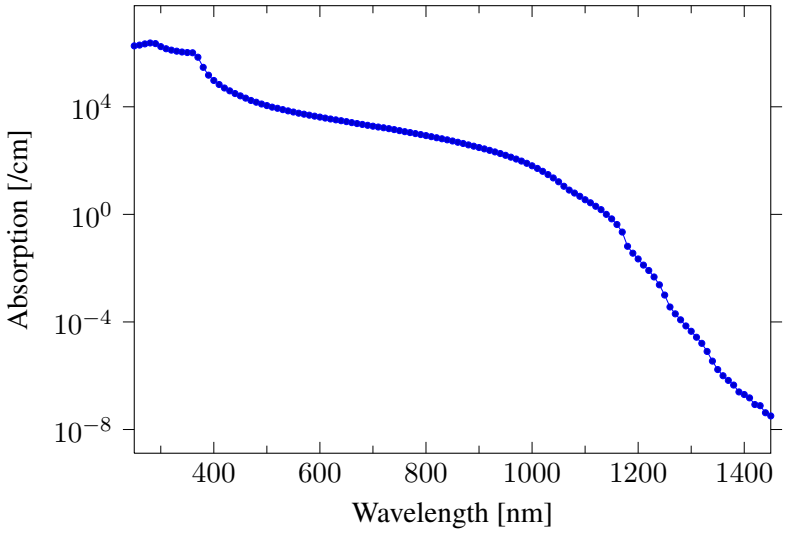


Figure 2.8: Silicon absorption coefficients. Source: [29]

Waveguides have been implemented in silicon and with the advent of Silicon-on-Insulator (SOI) technology highly integrated in a CMOS compliant substrate [30–32]. For example, figure 2.9 shows the cross section on a waveguide on an SOI wafer.

Silicon modulators built early in the 1990s by Treyz *et al.* [34] where leading the way to integrated light modulation. Since then, several SOI based modulators have been built [35–41]. The modulators use free-carrier absorption effects in p-i-n diodes combined with Mach-Zehnder interferometry. Recent modulators such those in [37–41] show modulation bandwidths in excess of 10 Gbps. An example of how light can be effectively coupled from a fiber into an SOI waveguide is shown in figure 2.10.

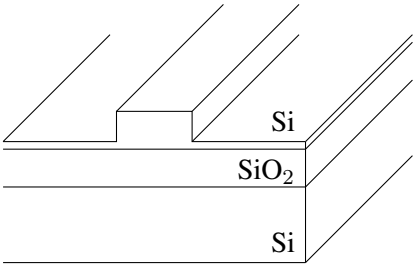


Figure 2.9: Cross section of a SOI based waveguide. For reference, the top silicon layer thickness can range from 12 nm to 100 µm, and the buried oxide layer thickness 10 nm to 3 µm. Source: [33].

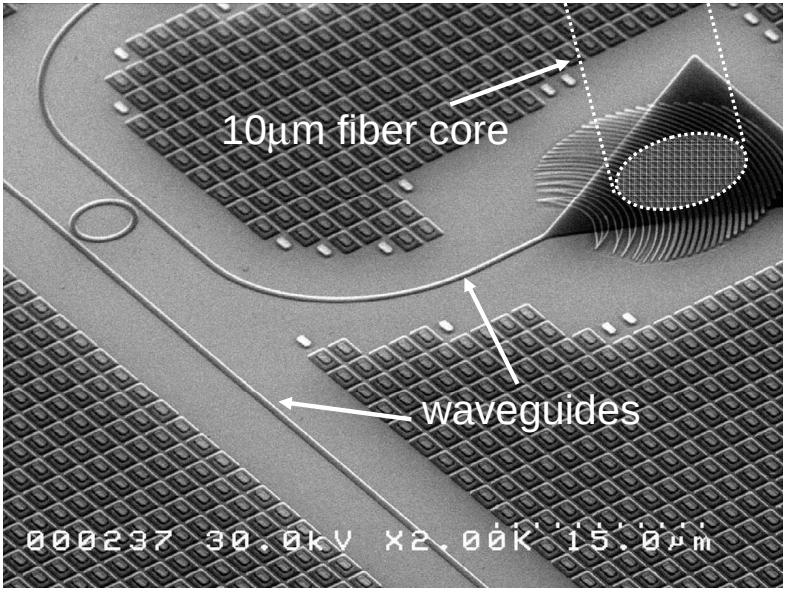


Figure 2.10: Example of fiber coupled into SOI waveguide with a holographic lens. Source: [37]

3 SINGLE-PHOTON COMMUNICATION

CHIP designers are developing increasingly complex integrated systems that require more and more die space for high throughput I/O pads. As a result, inter- and intra-chip communication is becoming one of the largest sources of noise and power dissipation on chip and also the bottleneck for performance. While transistor count has followed Moore's law; I/O pads have not evolved at the same pace. Moreover, due to bonding inductance, very high bit rates are possible at a cost of prohibitively high currents. Parallelism has often been used, but at a cost of large silicon area.

Traditional alternatives have been flip-chip and chip-level via technology [42]. However, reliability, cost, and flexibility are still open issues, especially when it comes to large inter-chip buses when more than two chips are involved. This is becoming an especially pressing problem with the emergence of densely packaged multiprocessor and multi-core systems. For this reason, 3D and system-in-package (SiP) techniques have been conceived to enable stacks of inter-bonded dies (see figure 3.1). The problem with this approach is however the limitation of speed imposed by bonding wires and the power dissipation of drivers.

To overcome these limitations, researchers have turned to wireless solutions based on capacitive, inductive, and optical methods [37, 43, 44]. While capacitive and inductive methods are effective in reducing power and ensuring high speed, they are only appropriate for pairs of chips. Hence, they are ineffective for broadcast and multi-chip systems.

Optical interconnects for inter-chip communication have been proposed decades ago. Their slow adoption is due mainly to the

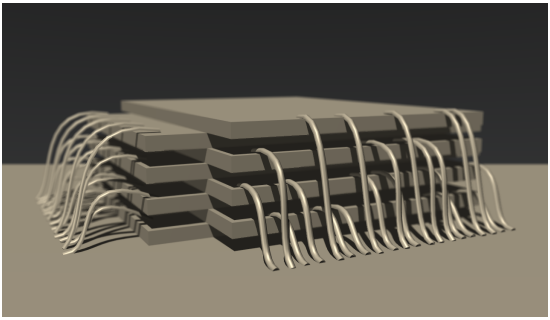


Figure 3.1: System-in-package wirebonding.

complexity of receivers whose output needs to be amplified, converted to a digital signal, and synchronized. These functions require area and may dissipate significant power. The lack of compact, low power optical sources has also been an issue. Commercial solutions for cabinet-level interconnect are actively developed by IBM, Intel, and Luxtera to name a few [2,3,45]. The current devices, generally built in CMOS SOI technology with Germanium deposited or flip-chip APDs, are not yet easily integrable into single-chip or multi-chip formats due to size and complex building processes although recent progress seem to address these issues [46].

In this chapter a new approach to optical communication is proposed that can be integrated in standard CMOS technologies utilizing a fraction of the area and power of a pad. The proposed approach is amenable to optical broadcasting, optical clock distribution, and optical buses (both vertical and horizontal). The core of the optical interconnect channel is a CMOS SPAD. Thanks to its digital output it requires no amplification, no A/D conversion, nor any other type of signal processing. However, in SPADs the detection cycle can be as high as a few tens of nanoseconds. Thus, a simple digital modulation scheme must be added to achieve throughputs of several gigabit-per-second.

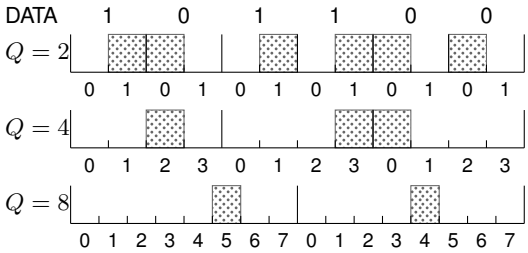


Figure 3.2: PPM modulation. The data to be modulated is shown at the top of the diagram. Q represents the length of the modulation slot. The position of the pulse in a slot represents the data.

We selected Pulse Position Modulation (PPM), a scheme that encodes K bits into $Q = 2^K$ time slots in the total allotted range $R = Q\Delta T$ (see fig. 3.2). PPM was selected to minimize the effects of SPAD dead time. In fact SPAD dead time in the order of a few tens of nanosecond, would limit On-Off Keying (OOK) modulation throughput to a few megabit per second. With PPM, we take advantage of the SPAD's photon-to-avalanche timing response whose jitter can be reduced to a few picoseconds. In fact, note that R should be higher than the detection cycle to ensure proper operation of the communication link.

A highly integrated application of the proposed system is depicted in figure 3.3. Optical data signals are generated, for example, in an integrated CPU by a micro LED similar to [18]. A sub-nanosecond optical pulse was demonstrated for this device using CMOS drivers that occupy a fraction of the area of a pad. Light is focused on the active region of a SPAD by micro lenses. Note also that optical chip-to-board and optical chip-to-chip communications could also benefit from this approach replacing optical channels with conventional fibers or waveguides for example.

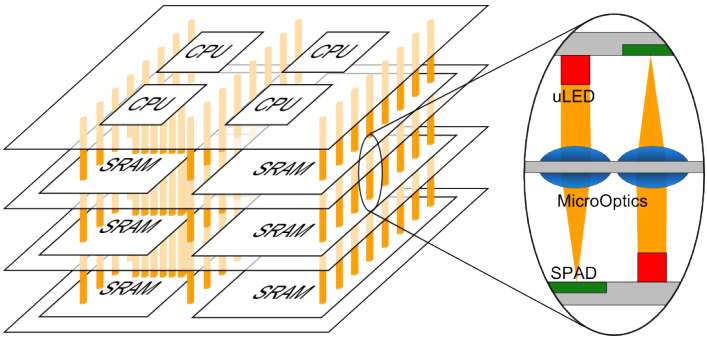


Figure 3.3: Inter-chip communication with LEDs, micro-lenses, and SPADs.

The detecting section of the channel is represented by a SPAD and integrated PPM decoder. The optical channel may be using integrated micro-optics that can be integrated on chip as a standard issue in most CMOS technologies. Multi-chip vertical buses can be obtained in this way by stacking dies that have been thinned. Optical transmission is ensured by low absorption coefficients of silicon in the visible spectrum.

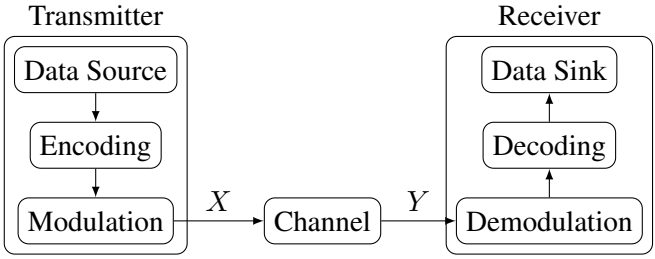


Figure 3.4: Point-to-point communication model.

3.1 Building Blocks

A traditional point-to-point communication system comprises the following elements: data source, encoder, modulator, channel, demodulator, decoder, and data sink. Figure 3.4 shows the chain relation between these elements. Several layers of protocol can be added on top of the data source/sink to provide reliability and flow-control. However, we will focus particularly on the modulation and demodulation part.

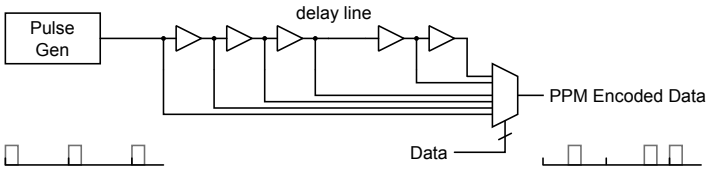


Figure 3.5: Simplified PPM modulation scheme.

Modulation of the data pulse in the proposed scheme can be implemented in several ways. Figure 3.5 presents a simplistic schematic of a modulator. The most critical issue in any modulator is to output pulses at a precise timing with a resolution ΔT that pushes the limits of the technology. The use of one or several delay lines (whether electrical or optical) seems a natural choice. Performing PPM modulation requires precise and fast timing control of light emission. Laser emission is generally best controlled in either constant emission or pulsed at fixed frequency. Shutter-based techniques combined with constant light source were proposed in [47] but bandwidth is limited by the shutter speed. A similar approach is to use interferometry as means of shutter to expand the bandwidth [40]. Pulse emission at a fixed frequency can be modulated with several delay lines multiplexed at the transmission rate [48–50]. Integrated light emitting diode such as [18, 22, 51, 52]

seem also good candidates for direct modulation. In particular, references [22, 51, 52] show sub-nanosecond optical pulses with an array of micro-LEDs.

Starting from the principle that optical to electrical conversion is available, demodulation of PPM signaling requires timing differentiation. The approach taken in [53] is to delay the integrated signal in Q delay lines, a pair-wise comparison of all signals is performed and fed into a "greatest of" receiver to recover the data. Shalaby in [48] counts photons in each interval ΔT and the interval with the maximum count is selected. Shalaby also proposes parallel sampling of the Q timeslots. Note that using SPADs it is difficult to count photons in an interval ΔT when ΔT approaches or is smaller of the SPADs dead time. However we can use the fast timing response of the SPAD to retreive the position of the first pulse in a frame. In this work, similar to [54], we propose to use a time discriminator such as a Time-to-Digital Converter (TDC) to retrieve the timing of the first pulse, hence the data. A TDC implementation is presented in section 3.3 with experimental results.

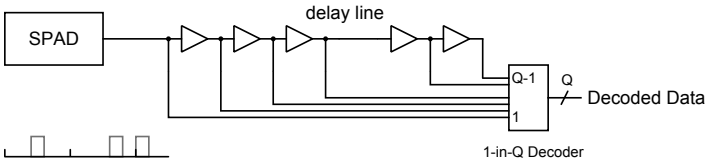


Figure 3.6: Example of PPM demodulation scheme.

Source coding is used to provide error detection and correction. Furthermore, it can be used to approach channel capacity [55]. Reed-Solomon codes and Turbo Codes are widely used for this purpose [56, 57]. Both are of particular interest when used with optical PPM communication as shown in [58, 59] for the former and in [49, 50] for the latter.

3.2 Theory and Limits

Optical communication channels have been extensively studied from early in the 70s to well after the 90s [47–50, 53, 58–66]. In this section, the results found in literature are re-derived with the SPAD peculiarities in mind.

3.2.1 Time Resolution Limits

Theoritical limits on the time resolution in an optical communication system were pointed out by Butman *et al.* in [62]. The fundamental limit is imposed by the quantum mechanics energy-time uncertainty principle:

$$\Delta E \Delta T \geq h, \tag{3.1}$$

where ΔE is the energy change in the system. The time resolution with which we can measure this change is represented by ΔT and h is Planck's constant. In other words, the time of the arrival of an energy changing event cannot be measured with arbitrary high precision. In the case of photons, the energy of a photon at frequency ν is $h\nu$ and equation 3.1 leads to

$$\Delta T \geq \frac{1}{\nu}. \tag{3.2}$$

At visible frequencies the resolution approaches 1×10^{-15} s or 1 fs. Note that while we detect photo-induced electrons and that timing of these events could be performed with arbitrarily high precision, the actual limitation comes from the photo-optical processing [62].

3.2.2 Channel Capacity

Butman in [62] already derived the bandwidth limitations in the case of a noiseless PPM-encoded optical channel. As also described in [63], the capacity of such channel can be made infinite, by decreasing ΔT. However, as derived in section 3.2.1, ΔT is ultimately limited by the quantum mechanic uncertainty principle. We will therefore derive the channel capacity by varying ΔT down to the 1 fs limit. In this section, we will first derive the channel capacity in the ideal case. We will then introduce SPAD's non-idealities and derive the channel capacity again.

The traditional idea of a point-to-point communication channel comprises a transmitter, a channel, and a receiver (figures 3.4 and 3.7). X and Y are random variables representing the transmitted codes before and after the channel respectively. The modulation proposed, PPM, uses Q time slots of ΔT seconds to encode Q source codes. A pulse present in time slot t_n represents the n$^{\text{th}}$ code.

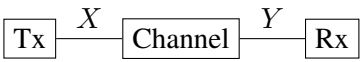


Figure 3.7: Point-to-point communication model.

The channel is modeled in the context of PPM modulation as a Q-ary erasure channel. In this model, a code is either received (not-erased) with probability $1 - \epsilon$ or not received (erased) with probability ϵ. The channel capacity as defined by information theory is determined by

$$C = \max_{P_X} I(X;Y), \tag{3.3}$$

where the mutual information $I(X;Y) = H(X) - H(X|Y)$. The

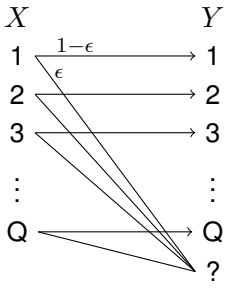


Figure 3.8: Q-ary erasure channel.

conditional entropy of X given Y is defined by:

$$H(X|Y) = \sum_y P(y)H(X|Y = y). \tag{3.4}$$

In the case of Q-ary erasure channel, the only non-zero term of the sum in eq. 3.4 is when $y = ?$. In fact when $y = \{1..Q\}$, knowing y completely determines x. The entropy

$$H(X|Y = ?) = H(X)$$

because knowing y does not give any information on x. From equations 3.3 and 3.4 the channel capacity:

$$\begin{aligned} C &= \max_{P_X} \big(H(X) - \epsilon H(X)\big) \\ &= \max_{P_X} (1 - \epsilon)H(X). \end{aligned} \tag{3.5}$$

The entropy $H(X)$ is maximized when X is uniform i.e.

$$\begin{aligned} H(x) &= -\sum_y P(y)\log_2 P(y) \\ &= Q \times \frac{1}{Q}\log_2 Q \\ &= \log_2 Q. \end{aligned} \tag{3.6}$$

Therefore the channel capacity of a PPM encoded optical channel is

$$C_{\text{ch}} = (1-\epsilon)\log_2 Q \qquad \text{bits/channel use.} \tag{3.7}$$

A channel use takes $Q\Delta T$ seconds therefore the capacity is

$$C = \frac{(1-\epsilon)}{\Delta T}\frac{\log_2 Q}{Q} \qquad \text{bits/s.} \tag{3.8}$$

In the noiseless case the erasure probability ϵ derives from the stochastic nature of light. Let us take a time interval ΔT and a light source of intensity Φ_{s} photons per second. The probability that n photons are generated in a time period of ΔT seconds is Poissonian with $\lambda = \Phi_{\text{s}}\Delta T$, hence:

$$P(n \text{ photons generated}) = \frac{(\Phi_{\text{s}}\Delta T)^n e^{-\Phi_{\text{s}}\Delta T}}{n!}. \tag{3.9}$$

The erasure probability ϵ is the probability that no photons are generated during period ΔT therefore

$$\epsilon_{\text{ideal}} = P(\text{no photons generated}) = e^{-\Phi_{\text{s}}\Delta T}. \tag{3.10}$$

The ideal channel capacity is plotted in figure 3.9 varying Q and ΔT. We note that 1 Gbps could already be achieved at a resolution of $\Delta T = 100\,\text{ps}$.

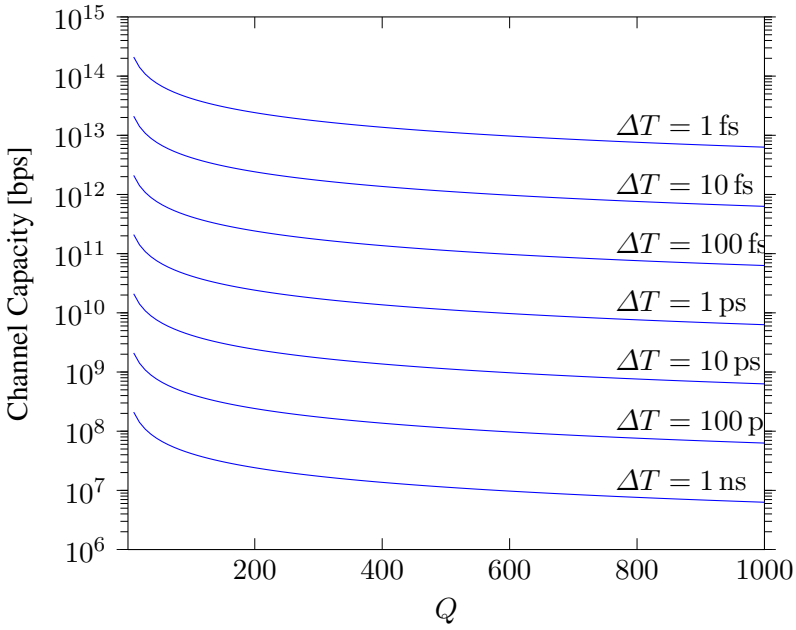


Figure 3.9: Ideal channel capacity with $\Phi_\mathrm{s}\Delta T = 1$.

When we take into account SPADs photon detection probability p_{pd}, the channel model described in the preceding paragraphs still holds. For simplicity we assume that the photon detection nonideality of the SPAD is in the channel and the receiver is ideal. We compute the erasure probability ϵ_{pdp} as the cumulated probability that no photons are detected assuming that photon detections in SPAD are independent and that the photon detection process is independent of the photon generation process.

$$\epsilon_{\mathrm{pdp}} = \sum_{k=0}^{\infty} P(k \text{ photons generated})(1-p_{\mathrm{pd}})^k \tag{3.11}$$

$$= e^{-\Phi_{\mathrm{s}}\Delta T} \sum_{k=0}^{\infty} \frac{(\Phi_{\mathrm{s}}\Delta T(1-p_{\mathrm{pd}}))^k}{k!} \tag{3.12}$$

$$= e^{-\Phi_{\mathrm{s}}\Delta T} e^{\Phi_{\mathrm{s}}\Delta T(1-p_{\mathrm{pd}})} \tag{3.13}$$

$$= e^{-\Phi_{\mathrm{s}}\Delta T p_{\mathrm{pd}}}. \tag{3.14}$$

Equation 3.11 is equivalent to saying that the probability of erasure is the sum of the probabilities that for the period of time ΔT, zero photons are generated, one photon is generated and one photon is not detected, two photons are generated and two photons are not detected, etc. In (3.12) we use the well known series expansion $e^x = \sum_{k=0}^{\infty} x^k/k!$. In figure 3.10 the channel capacity is plotted against PDP given $\Phi_{\mathrm{s}}\Delta T = 1$ and $Q = 2^{\dagger}$. A PDP of 10 % yields a drop in channel capacity of one order of magnitude.

[†]For such low value of Q and ΔT, a very low SPAD dead time is required for proper operation.

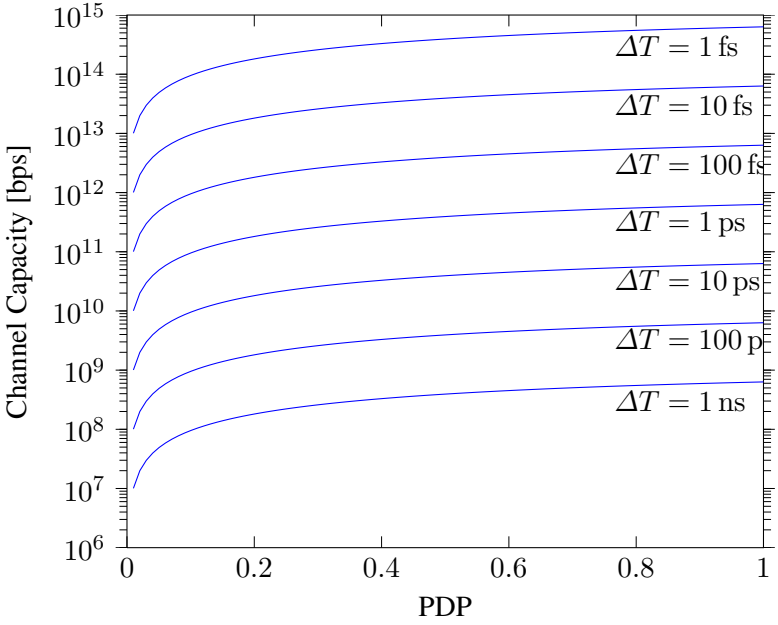


Figure 3.10: Channel capacity variation according to photon detection probability. $\Phi_s \Delta T = 1$, $Q = 2$.

In order to derive the capacity in presence of noise, we need to change the channel model. Figure 3.11 shows the mapping of the chosen Q-ary noisy erasure channel. The conditional distribution $P(X|Y)$ completely defines the channel transition probabilities hence the capacity. The decoding strategy, which influences $P(X|Y)$, is defined as follow. For each period $Q\Delta T$, the timing of the first pulse defines the code. The noise is modeled as a Poisson process with rate λ_N. We define N to be the time of arrival of the first noise pulse (we assume noise pulses are not erased). The random variable N has an exponential distribution with rate λ_N. The conditional distribution $P(X|Y)$ for the Q-ary noisy erasure channel is

$$P(Y=j|X=i) = \begin{cases} \epsilon_j - \epsilon_{j+1} & j < i \\ \epsilon_j(1-\epsilon_{\text{pdp}}) + (\epsilon_j - \epsilon_{j+1})\epsilon_{\text{pdp}} & j = i \\ (\epsilon_j - \epsilon_{j+1})\epsilon_{\text{pdp}} & j > i \\ \epsilon_Q \epsilon_{\text{pdp}} & j = ? \end{cases} \tag{3.15}$$

where $\epsilon_i = P(N > i\Delta T) = e^{-\lambda_N i \Delta T}$.

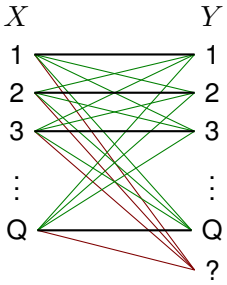


Figure 3.11: Q-ary noisy erasure channel.

The solution of equation 3.3 in the general case requires to find the distribution of X that maximizes the mutual information $I(X;Y)$. Equation 3.3 can be rewritten as

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} \sum_x \sum_y p(x)p(y|x) \log \frac{p(x|y)}{p(x)}. \end{aligned} \tag{3.16}$$

While equation 3.16 can be solved algebraically in several cases, we chose to use Arimoto-Blahut algorithm [67, 68] to solve it numerically with $p(y|x)$ given by equation 3.15. The capacity per channel use was computed with noise ranging from a few herzt to one gigahertz. The results in figure 3.12 show that even in the case of noise as high as a few kilohertz, channel capacity is very close to the noiseless case.

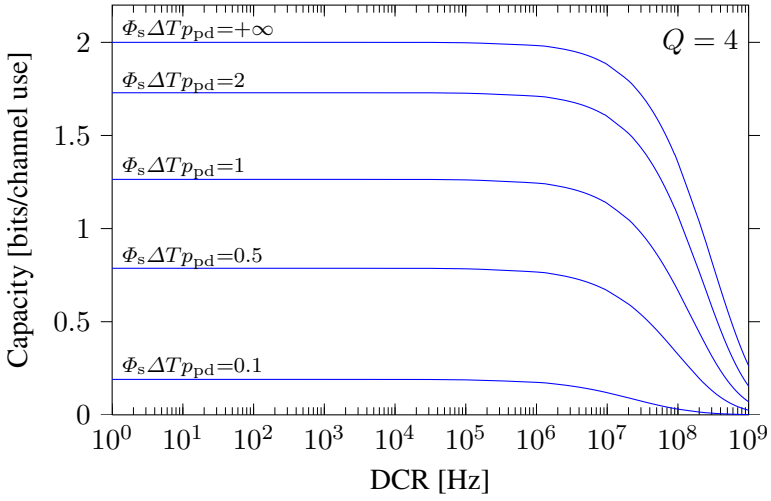


Figure 3.12: Capacity per channel use in presence of noise and erasure. $Q = 4$.

SPAD's dead time and afterpulsing should also be addressed. The effects of dead time can be mitigated in several ways. Let τ_{dt} denote the dead time. Imposing $\Delta T > \tau_{\mathrm{dt}}$ would be very limiting in bandwidth unless τ_{dt} can be made arbitrarily small. Relaxing the requirement to $Q\Delta T > \tau_{\mathrm{dt}}$ enables large channel capacity but, as the channel looses its memoryless property, state-dependent coding must be performed. Afterpulsing is highly linked with dead time. It can be seen as data-correlated noise which further reduce channel capacity. Source coding, such as block length coding, could be used to alleviate some of the these effects, however, we didn't yet expand our research in this area beyond the following paragraph.

Although we derived the channel capacity as a function of several system design variables, care should be taken in analyzing the numbers. As channel capacity is an upper limit to the transmission rate achievable, the practical limits are actually that of the source coding schemes implemented. We will not discuss source coding besides the fact that well known coding schemes such as Turbo Codes and Low Density Parity Check (LDPC) codes have shown performance approaching channel capacity. Area, power, and performance of such a system need to be analyzed in a prototype in order to finish this discussion. A first step toward this goal is presented in the next section, where we review time discrimination.

3.3 Time discrimination

This section is devoted to the measurement of time, that can be applied, but not limited to, PPM demodulation. In many applications, the precise measurement of the time difference between two signals spaced in time is important, especially when digital pulses are used relating to different physical measurements. Examples of such applications include time-of-flight sensors, optical correla-

tion spectroscopes, positron emission tomography instruments, and pulse position demodulators [69–71]. High time resolution, typically 100 ps, is key to achieving the desired performance, while high throughput is often required as well. For instance, in metrological applications, high throughput has the effect of achieving the wanted accuracy in a shorter time, thus improving the speed of a measurement or making it possible to achieve unprecedented accuracies whenever a phenomenon is fast occurring. Similarly, apparatuses, such as LIDARs, LADARs, and RADARs, as well as techniques using time-resolved imaging can benefit from this functionality [72, 73].

We look to implement time measurement with Field Programmable Gate Array (FPGA) devices. However some of this discussion also apply to custom ASIC designs. Several techniques exist to measure a time difference; the simplest is based on a counter. In principle, the clock frequency of the counter determines the resolution. Thus, if metastability effects are ignored, to achieve 100 ps of resolution, a 10 GHz clock is required. Due to bandwidth limitations, it is generally not feasible to achieve such clock speeds in most commercial FPGAs. Alternatively, a time-to-amplitude converter (TAC) can be used. TACs enable detection of extremely short time differences by translating time onto voltage that can then be measured with high precision. However, this approach requires high-speed analog switches, high-precision current sources, and, in most cases, high-resolution A/D converters, components generally not available in conventional FPGAs.

An alternative is the use of a time-to-digital converter (TDC). TDCs are digital components that can be implemented in a fully-digital design style either as an ASIC or an FPGA. The literature on this type of devices is quite extensive, however, to our knowledge, none of them have implemented simultaneously in FPGA sub-nanosecond resolution *and* high throughput [74–85]. High time

resolution and high throughput are often contradictory specifications, however by proper use of design techniques and heavy use of pipelining, it is generally possible to achieve a good compromise. In the design proposed in [86] for example, a 3-stage pipelined TDC was used to achieve 10 MS/s with a resolution of 97 ps. However, the TDC was designed in full-custom style and it occupied a surface of approximately 0.4 mm^2. In addition, the design was highly optimized, so as to achieve Process, Voltage and/or Temperature (PVT) variability control via a built-in feedback loop.

This section develops the FPGA-based TDC architecture presented in [83, 87, 88]. Two implementations of the architecture are presented. For the purpose of clarity, the latest implementation in a 65 nm FPGA is presented and the 130 nm FPGA implementation is shortly discussed in section 3.3.5. The 65 nm Virtex 5 implementations achieves an overall resolution of 17 ps over a range of temperatures. The TDC requires only 1208 slices while the throughput is 20 MS/s in normal operating mode and 300 MS/s in Turbo mode. The TDC exploits several design techniques to best utilize the available technology while maintaining high utilization efficiency. A repetitive fabric of CLBs is thus utilized to minimize clock net distribution mismatches while limiting skew. In addition, two calibration techniques are proposed for countering PVT variations albeit focusing mainly on temperature compensation.

The PVT variability control can be achieved either via a power supply control feedback on board or using an entirely digital scheme. The latter technique is based on the use of a periodical measurement of the fine delay of the TDC's built-in delay line. In case of temperature variations, each delay element will change its propagation delay, thus the number of delay elements will be adjusted. To maintain virtually the same resolution, the raw codes are mapped to the correct time difference using an appropriate table, while inter-value interpolations may be used. The same technique can be used when

migrating the TDC configuration from one FPGA to another and from one power supply to another, so as to achieve PVT independence.

3.3.1 Architecture

General Considerations

Several time resolution methods are usually used in order to extend the range of measurement of time-to-digital converters. The Nutt interpolation technique splits the measurement between a coarse and fine (figure 3.13). At the coarsest level, a counter is usually employed, achieving nanosecond resolutions. At finer levels, the two most used approaches are

1. The Vernier method: it reaches its finest resolutions by harnessing the difference of propagation delays of the elements used. Generally two oscillators slightly out of tune [89,90] or two tapped delay lines with slightly different delay [91, 92] are used.

2. The tapped delay line method: the method uses compensated [86, 93–96] or non-compensated [76–78] tapped delay lines. The resolution here is inherent to the delay element used.

These techniques have been mapped to FPGA with relative ease. References [74–76, 81] use the Vernier method with mismatched delay lines while [80,85] use ring oscillators. In [77–79,82,83], like in this thesis the tapped delay line method based on carry propagation lines is employed. Dedicated carry lines present nowadays in all FPGA fabrics are usually employed in adder logic and therefore each adder's bit has a flip-flop nearby. For example, the Virtex-5 slices carry logic is shown in figure 3.14.

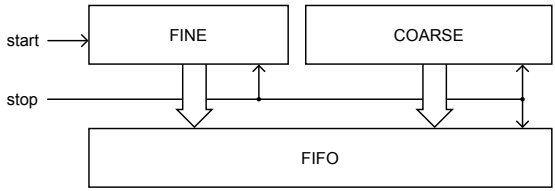


Figure 3.13: The global architecture of the TDC uses the Nutt interpolation technique whereby the measurement is split between coarse and fine.

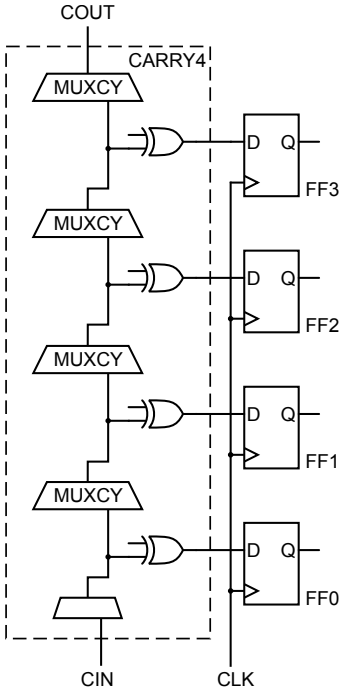


Figure 3.14: Xilinx Virtex-5 slice (partial) with carry logic and sequential logic. CIN and COUT are inter-slice carry signal ports from the slice below and to the slice above respectively.

The architecture of the presented TDC, as shown in figure 3.15, consists of a free running coarse counter, a tapped delay line implemented in a carry chain, an input signal filter, an encoder, reset logic, and readout logic. The principle of operation is the following. Time is split in frames of 16 coarse clock cycles. The state of the delay line is latched at every clock cycle. A three-stage synchronizer is used to minimize metastability effects. Whenever the hit signal arrives, it propagates in the delay line. The detection and measurement is then performed by detecting transition in the first bits of the line. The thermometer encoded value in the line is converted to binary. Finally, the coarse and fine values are read out through a FIFO. A clock cycle is reserved for resetting the state of the delay line and filter, therefore one single measurement may be performed per frame.

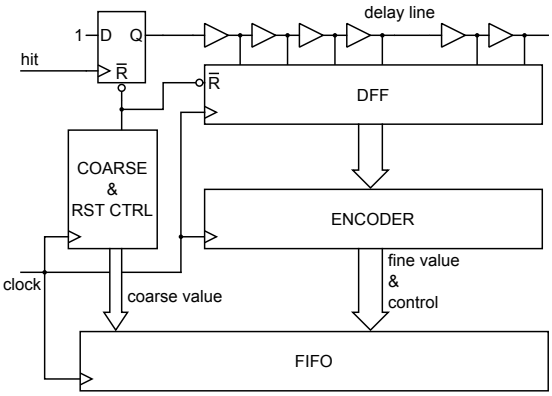


Figure 3.15: Normal mode architecture of the TDC. The hit signal is fed to the delay line through a filter flipflop. The state of the delay line is latched at the rising edge of the clock which acts as stop signal. The reset circuitry, coarse counter, encoder, and FIFO are in the same synchronous domain.

Setup time or hold time of the fine chain latches are sometimes violated in the normal operation of the TDC. The value of violating latches will eventually resolve into a one or a zero. However, this can produce *bubbles* in the code, even when assuming a perfectly unskewed clock distribution. In the presence of clock skew, this phenomenon also occurs. A bubble-tolerant thermometer-to-binary encoder is required to limit the effect of bubbles on the output codes.

The design was implemented in VHDL with a completely automatic place and route procedure. Constraints were used to place the delay chain in the desired locations and no manual routing was performed.

Reducing Dead Time

The hit signal propagates from the filtering flipflop to the fine delay line. Assuming an all-ones reset state of the line, zeros propagate from the beginning of the line toward the end. Since the line is sampled at each clock cycle, the resulting succession of *fine states* could be (bits flowing right to left):

coarse	fine line
0	1111111111111111111
1	1111111111111100000
2	1110000000000000000
3	0000000000000000000
...	...
14	0000000000000000000
15	1111111111111111111

The measurement is armed at cycle 1 while cycle 15 resets the line to start a new measurement. The *dead time* is defined as the minimum time between the end of a measurement and the start of

the next one. For the architecture shown in figure 3.15, the dead time is one clock cycle. To reduce it, we propose to make use of propagation of ones and zeros alternatively with a toggling filter (fig. 3.16). With this architecture, the dead time is reduced to the minimum toggling time of the input filter flip-flop (<500 ps).

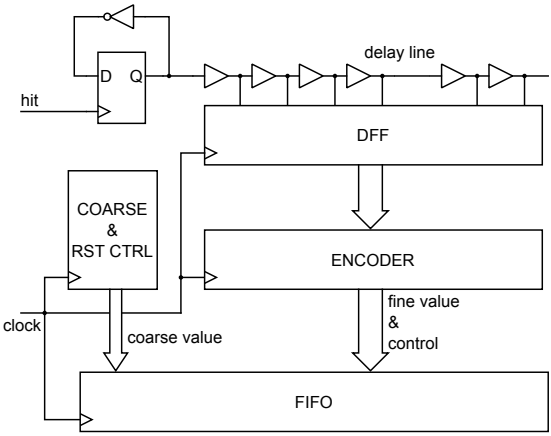


Figure 3.16: Turbo mode architecture. The two main differences from the normal mode architecture are: 1. the hit signal is filtered with a toggling flip-flop, hence '0's and '1's propagate through the delay line alternatively. 2. The filter and the line state latches are not reset anymore.

This architecture, referred to in the remainder of this chapter as Turbo mode, has the advantage of increased readout speed and the ability to handle multiple hits per frame. There are three main limitations of this design. First, the high-to-low and low-to-high transitions formed by the traveling hit signal have different propagation delays in the delay line. This yields an asymmetry in the measurement. Second, the metastability of the delay line latches is higher than in the normal architecture. The latches and input filter are never reset, therefore a metastable state can persist depending

on the clock frequency used. Third, the complexity of the encoder increases if detection of multi hits per clock cycle is required. Note however that our encoder can only encode a single measurement per clock cycle.

3.3.2 Digital Calibration

As traditional compensation techniques are not readily available in FPGAs, we need to rely on digital techniques to calibrate PVT variability. A method for mapping code to real-time value is necessary and of particular importance is the range of elements used in the fine line from which is derived the resolution. Several techniques to address these issues are presented in this section.

Interpolation

Let N_{max} be the total number of elements in the chain. The interpolation process is the procedure by which a code c is converted from its natural interval $(]0, N]$, with $N < N_{max}$) to an output interval $(]0, 2^b])$. Linear interpolation as the simplest approximation is defined:

$$c' = \left\lceil \frac{c \cdot 2^b}{N} \right\rceil$$

Linear interpolation can be conveniently implemented through mapping. The map information stored in a RAM block can be generated on-the-fly whenever the range changes. A single RAM lookup can output the interpolation result. As shown in figure 3.17, some codes may never appear in the output of the mapping if $2^b > N$ and two or more input codes can be mapped to the same output code if $2^b < N$.

The real bin width of the fine elements can be estimated through a statistical code density test [97]. The interpolation process can be

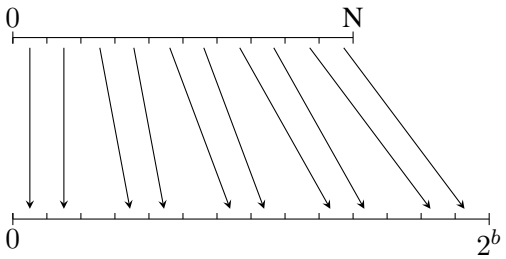


Figure 3.17: With linear interpolation, output codes are skipped so that the N input codes are mapped into the $]0, 2^b]$ range.

improved with this knowledge. If the output code space is chosen such that $2^b > N$, then there exists a source bin (a *large* bin) that is mapped to several destination bins. A pseudorandom distribution of original bin values to the corresponding output bins can be performed. For example, in figure 3.18, given the statistical code density test result, we can build an intermediate representation of the binning interval. This interval is the real-value time interval of one clock cycle where each bin has its *size* set according the density test. The interval is quantized into 2^b slots. A given bin will then map to the corresponding slots to the extent of the bin coverage. For example the last bin in figure 3.18 will map to output codes $2^b - 1$ and $2^b - 2$ with likelihood of 40 % each and to $2^b - 3$ with likelihood of 20 %. This non-deterministic mapping can be implemented on FPGA with a pseudorandom number generator or as a post-processing step. We call this method *pseudorandom bin dithering* in the remainder of this chapter.

Automatic Range Adjustment (ARA)

Let $N(x)$ be the number of propagated bits in the fine delay line latched at cycle x. Measurement $N(x)$ is valid if $N(x) > 0$ and

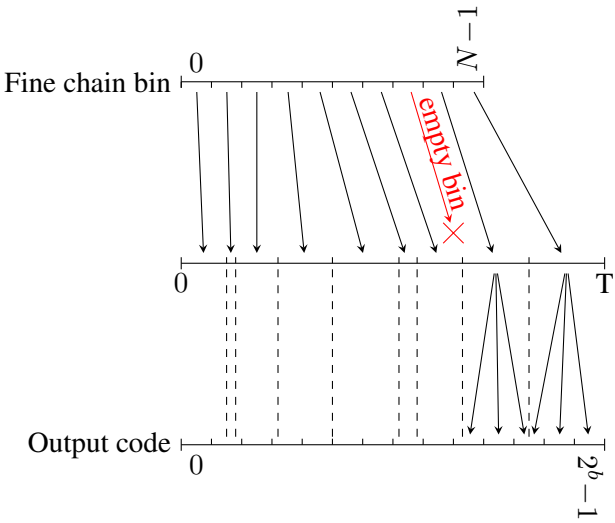


Figure 3.18: Pseudorandom bin dithering based on statistical code density test results. The statistical code density test is used to infer the real bin width of the fine chain bins. Each bin is first mapped to the real-value time interval spanning one clock cycle. Then, for each output code a probability density function is computed. In the example, the fine chain bin $N-3$ was found empty after the code density test, hence not output code space is allocated.

$N(x-1)=0$. Measurement $N(x+1)$ is a range measurement if $N(x)$ is valid and $N(x+1) \neq N_{max}$. The average resolution of the fine delay line r or 1 LSB is defined by

$$r = \frac{T}{N(x+1) - N(x)}$$

where T is the clock period. In the example of table 3.1, $x+1$ is a range measurement and therefore $r = T/10$.

coarse	fine line	$N(x)$
$x-1$	11111111111111111111	0
x	11111111111111000000	6
$x+1$	11110000000000000000	16
	|---------|	

Table 3.1: Automatic Range Adjustment Example: measurement at cycle x is valid. measurement at cycle $x+1$ is a range measurement. The resolution $r = T/10$

This method has the advantage of enabling to adjust the range on-the-fly while measurements are performed. A running mean of the range value can be kept and the interpolation process can be adapted consequently. However, this method can only adapt to relatively slow changing variations in PVT (in the order of several hundred clock cycles) and requires that some of the measurements yield a range measurement. This last requirement might never happen, therefore a simple range calibration scheduling can be put in place in order to gather these periodically.

Downsampling

The inter-slice routing delays are expected to be relatively large compared to those inside a slice. Better uniformity can be obtained by implementing only 1 tap per slice at the cost of reduced resolution. While we decided to use the maximum resolution the fine line can offer by implementing 4 taps per slice (see Fig. 3.14), some results of the downsampling technique are presented in section 3.3.3.

3.3.3 Results

Test Setup

A Xilinx ML505 board was used to perform our experiments. The on-board Virtex-5 device is fabricated in 1 V, 65 nm triple-oxide process. High frequency SMA connectors were employed to feed the *hit* signal to the FPGA clock inputs. A custom USB interface board is used to transfer measurements to a computer. All reference measurements, such as input clock jitter, cable time delay, and time distribution of the random input signal, were performed on a LeCroy WaveMaster 8600A digitizing oscilloscope.

The clock (*stop* signal) of our design is generated on board by a low jitter frequency synthesizer. The differential output of the synthesizer after on-board buffering and an FPGA pass-through connection shows a measured jitter of less than 12 ps at 300 MHz. The coarse counter is free running at this frequency and shows an integral linearity error below 0.03 coarse LSBs. Our focus for the remainder of this section will be on the fine interpolation technique.

Three types of measurement were performed: linearity characterization, ARA validation, and precision evaluation.

The linearity of the TDC can be characterized in several ways. A precise variable delay generator such as a tapped mechanical rail can be used to cover all of the fine TDC measurement range and derive the linearity. This yields stair shaped graphs such as those found in [74, 79, 98–101]. Another commonly used method is the statistical code density test [74]. The idea is to generate a large number N of pulses randomly distributed in time, and to collect the result of the TDC interpolation into a histogram. In the ideal case the histogram has C code bins containing $\overline{n} = N/C$ counts each. In reality there is a differential non-linearity $\mathrm{DNL_c}$ and for each

code $c \in [1, C]$, $n_c \neq N/C$, n_c representing the depth of bin c.

$$\mathrm{DNL_c} = \frac{\mathrm{n_c}}{\overline{\mathrm{n}}} - 1.$$

The cumulative sum of $\mathrm{DNL_c}$ yields the integral non-linearity $\mathrm{INL_c}$

$$\mathrm{INL_c} = \sum_{\mathrm{i}=1}^{\mathrm{c}} \mathrm{DNL_i}.$$

INL and DNL values refer to the 1-bin unit hence are expressed in LSB units. Note also that, in practice, a random pulse generator is not required as a non-stabilized oscillator suffice as long as the output is not correlated with the TDC's main clock.

The automatic range adjustment values are gathered whenever available. All the values are collected into a histogram, and we derive the mean bin width of the fine codes.

To measure the standard uncertainty or random error, we inject a delayed version of a clock synchronous signal into the fine delay line. Several fixed length cables are used to generate delays. Several hundred thousand measurements are taken and the worst standard deviation of the resulting codes is reported.

The pseudorandom bin dithering technique described in section 3.3.2 has also been evaluated. While its implementation in FPGA is conceivable, we post-processed gathered traces to produce the DNL and INL graphs shown.

Temperature measurements were performed to validate ARA and to evaluate the variability in temperature. The temperature was set from 30 °C to 80 °C in a kiln. The measurement were taken after a stabilization period of up to 15 minutes. Both ARA and INL variations are reported.

Measurements

The statistical code density tests were performed with a photodetector exposed to ambient light. The sensor based on a single photon avalanche diode (SPAD) [5] displays a uniform distribution of pulses separated by at least 30 nanoseconds. Note that the same test can also be performed with an oscillator non-correlated with the system clock. In fact, the phase variations, between the oscillator signal and the system clock, due to non-tuned frequencies and PVT noise can ensure uniform coverage of the time interval. Such oscillator could also be implemented directly in FPGA, allowing *in situ* calibration. Figure 3.19 shows the INL and DNL of the fine line. The obtained DNL varies between -1 and $+3.55$ LSB. The INL lies in $[-3, +2.58]$ LSB range.

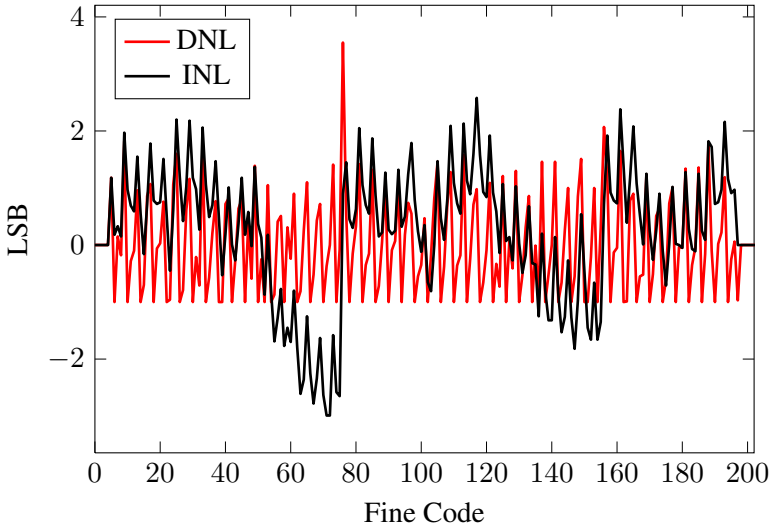


Figure 3.19: Performance of the fine measurement. Differential and Integral Nonlinearity.

The large variations of DNL in figure 3.19 around bin 76 are due to the clock skew between slices. The clock signal travels from the differential input clock pads to a global clock buffer (BUFG) and is then distributed through regional buffers to each slice. The delay from the BUFG to some slice is reported in table 3.2. Note that the difference between slice X28Y39 and X28Y40 is 57 ps (=3.3 LSB).

Slice	Delay	Fine position
X28Y21	1.614ns	3
X28Y22	1.612ns	7
X28Y23	1.609ns	11
	...	
X28Y37	1.613ns	67
X28Y38	1.615ns	71
X28Y39	**1.617ns**	75
X28Y40	**1.560ns**	79
X28Y41	1.558ns	83
X28Y42	1.556ns	87
X28Y43	1.553ns	91
X28Y44	1.549ns	95
X28Y45	1.544ns	99

Table 3.2: Clock distribution delays for fine TDC line.

The results of our statistical code density tests show that inside each slice there is a disparity between the four latches (fig. 3.14). Besides the fact that the distribution across the four latches is not uniform, the second latch in every slice never becomes the most significant propagation bit. This systematic error can be explained by the combination of two factors. First, the propagation delay inside the CARRY4 structure is not uniform. The simulation results shown in table 3.3 display a delay for bin 2 that is substantially

smaller. Second, the clock skew between the latches is similar to the inter-slice clock skew problem of the previous paragraph. These phenomena contribute to the creation of bubbles.

Point	Delay [ps]	Diff [ps]
CIN	0	0
FF0_D	33	33
FF1_D	47	**14**
FF2_D	81	34
FF3_D	104	23

Table 3.3: Carry4 structure delay from CIN to FF (from simulation).

Several placements of the delay chain were tested and this non-linearity was found in all the designs. With further inspection we noticed that the anomaly arises when the chain crosses the middle of the device from slice X*Y39 to slice X*Y40. Since our design needs at least 50 slices (200+ bins) to cover a clock period, this non-linearity cannot be mitigated with chain placement. The clock skew variability underlying this problem was described in detail in [102]. To be able to remove the problem without changing device, we propose two solutions. The first solution is to compensate the clock skew by generating delayed versions of the clock for each problematic region. This can be done either with DCM or IODELAY elements. Note however that DCMs inject significant jitter (see section 3.3.4). The second solution is thus recommended to shorten the fine line and place it in such a way to avoid the particularly bad spots. To reduce the delay line, either the clock frequency needs to be increased or the line must be split in two; it thus operates on both clock edges. This has to be done in order to always have a complete clock cycle covered.

Downsampling by 4 (1 bin per slice) leads to improved INL as shown in figure 3.20. Here the INL's range is $[-0.49, +1.18]$ LSB. The drawback of this approach is the loss of resolution. In this context, for about the same resolution as [78], the INL of our system $[-0.49, +1.18]$ LSB is narrower than in [78] $[-2.003, +1.855]$ LSB. While the clock skew problem was also reported in that work, note that the device used (Virtex II) is fabricated in a 130 nm, 1.5 V CMOS process.

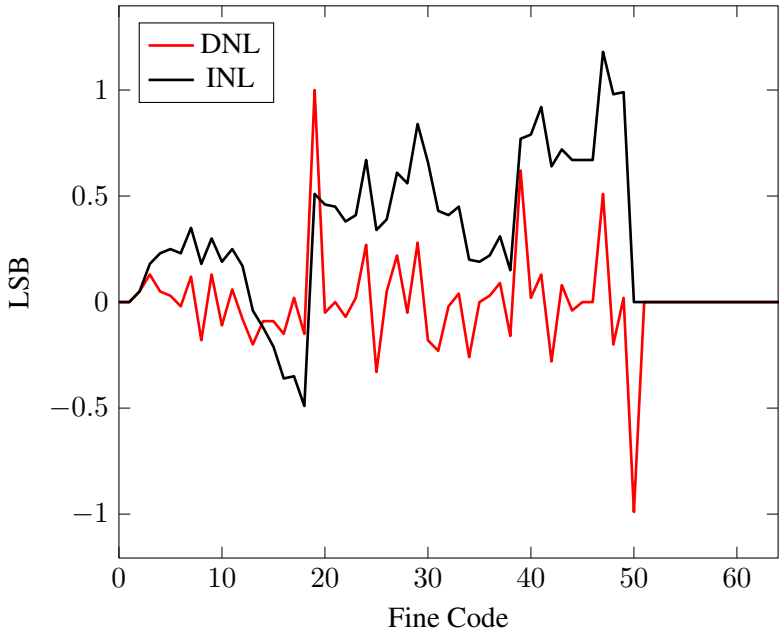


Figure 3.20: DNL and INL of the fine delay line after downsampling by 4 (1 tap per slice).

The collected values for the automatic range adjustment are shown in figure 3.21. The histogram of the values taken at 30 °C, yields to the conclusion that, on average, 207 elements are used to cover a clock cycle of 3.3 ns. The bin width is therefore, on average, 16.1 picoseconds. The standard deviation of the histogram is of about 2 bins.

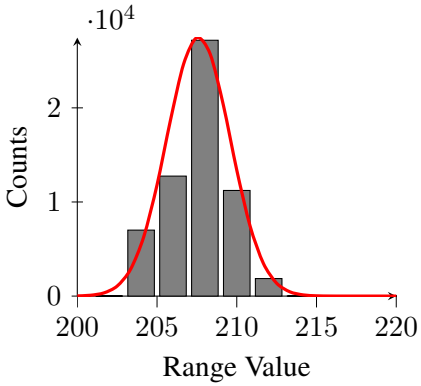


Figure 3.21: Histogram of the automatic range adjustment method.

The standard uncertainty was measured with fixed length cables. As expected the uncertainty or precision of the measurement degrades as more elements of the delay chain are used. The main cause being the accumulation of jitter while the signal passes through the buffers of the chain. The worst-case standard deviation is reported in table 3.4. For the example given in figure 3.22, its value is 20.46 ps.

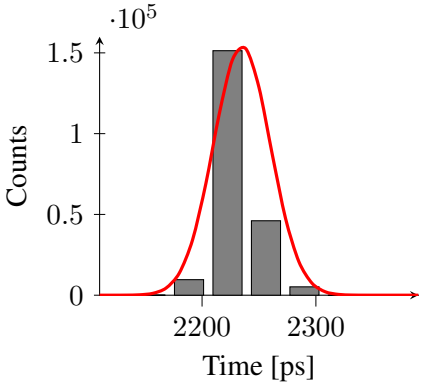


Figure 3.22: Time delay measurement, std dev: 20.46 ps, binning after downsampling by 2

The pseudorandom bin dithering technique described in section 3.3.2 was performed as follows. First, a reference code density test trace is taken and the bin width and mapping probabilities are computed. Then, a different trace is played back through the mapping. Finally the DNL and INL were computed. The resulting DNL and INL are shown in figure 3.23. The results shown were produced off-line in a post-processing step and the INL range is $[-0.27, +0.66]$ LSB.

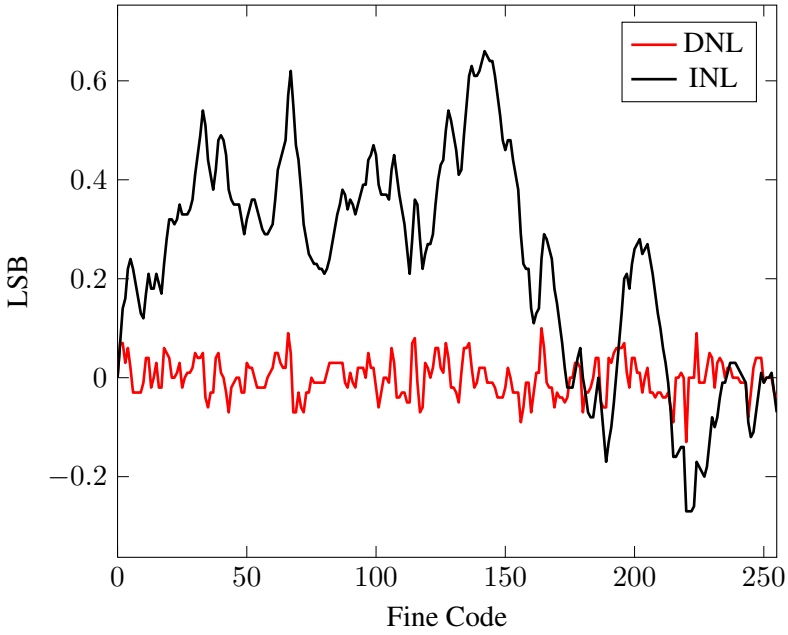


Figure 3.23: Results of dithering. First a sample code density test is taken as reference. Then, subsequent time measurements are converted to output codes with the fixed pseudorandom mapping.

While it may not increase precision or accuracy by itself, this technique has the advantage of mapping the resulting codes to a known range. The conversion from code to real time value is therefore facilitated. Note also that, due to the inherent construction of the algorithm, the INL and DNL presented can only be achieved on average. Note that when this requirement is not necessary, a linear mapping is sufficient.

The degradation of INL for a temperature variation from 30 °C to 80 °C is shown in figure 3.24. The TDC was calibrated at 30 °C and the INL was computed for the traces taken at different temperatures.

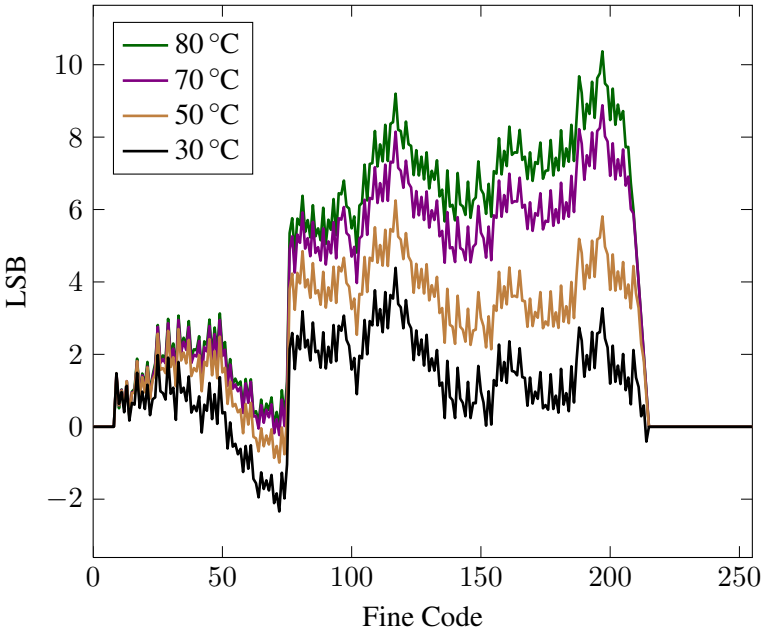


Figure 3.24: INL degradation in temperature. The TDC is calibrated at 30 °C and the INL is computed at 30 °C, 50 °C, 70 °C, and 80 °C.

The recorded variation on the ARA values is shown in figure 3.25. As expected the range of used elements in the fine delay line is reduced as temperature increases. The INL after correction was recomputed and is shown in figure 3.26. Note also that the resolution is impacted by the variation.

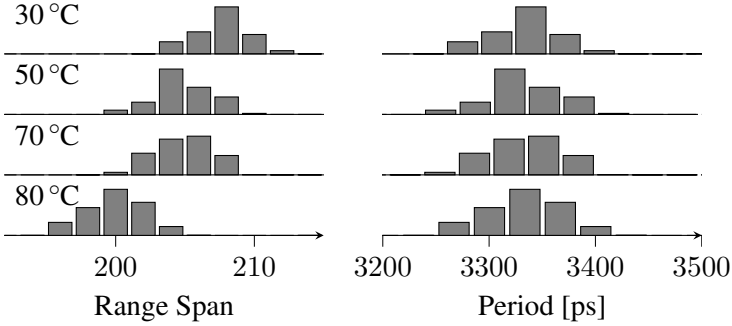


Figure 3.25: Left: ARA value variation in temperature. As expected the range of used elements in the fine delay line is reduced as temperature increases. Right: ARA values translated in time domain. The mean of the ARA values always represent the fine range of measurement i.e. the clock period of 3333 ps in our design.

Process, Voltage, and Temperature variability has a major impact on TDC design, especially in FPGAs where less design flexibility is available. Process variations come in two varieties: intra-die and die-to-die. Both are due to stochastic fabrication differences in doping levels and geometry. However, given a routed design, its characteristics are fixed and a simple calibration can compensate process variations. Voltage variations will affect the TDC and the fine delay line by slowing or accelerating propagation. The variation can be sudden and of varying amplitude. Therefore it is very hard, if not impossible, to compensate it by means of calibration.

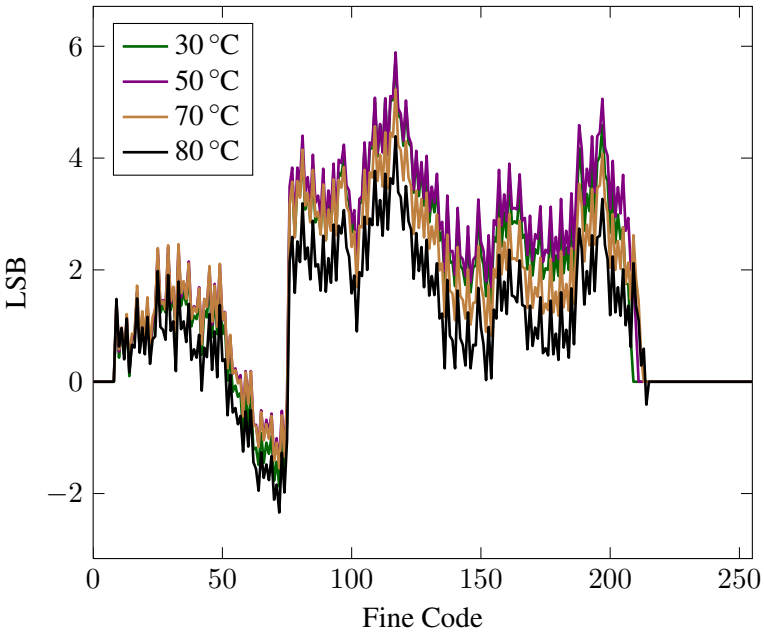


Figure 3.26: INL in temperature corrected with the ARA method.

Good decoupling practices should suffice to effectively reduce voltage variations, however care must be taken that the internal digital logic does not produce local IR drops which cannot be easily controlled. Finally temperature variations have similar repercussions, however, their rate of change in the order of several seconds, is rather slow. For this reason, a calibration as presented in this work (ARA) or other periodic calibration schemes can be effective. Note the ARA is an online mechanism that takes advantage of measurements that lie in the beginning part of the line and hence get to propagate for a full clock period in the delay line. For the same reason, ARA might not be completely accurate as the results show. In fact, the number of elements returned is measured with a ma-

jority of but not all the elements of chain that are used in a normal propagation. Probably, this is the reason behind the discrepancy between 50 °C and 70 °C ARA values, while the INL at 70 °C shows the worst non-linearity.

The Turbo mode architecture leads to results that are similar to the normal mode. The mismatch between the propagation of zeros and ones is of $2 - 3$ bins. We were not able to see metastability errors in the codes. However, in our experiments, increasing the main clock frequency with this architecture leads to metastability.

Table 3.4 summarizes the characteristics of this work.

	Min	Typ	Max	Unit
Clock frequency		300		MHz
Standard uncertainty	9.8		24.2	ps
Resolution		16.9		ps
DNL	-1		3.55	LSB
INL	-2.99		2.58	LSB
Pseudo-random bin dithering				
DNL	-0.13		0.1	LSB
INL	-0.27		0.66	LSB
Normal Mode				
Measurement Range (MR)		50		ns
Dead Time (DT)	3.33		50	ns
Readout speed		20		MSample/s
Turbo Mode				
Measurement Range (MR)		53.33		ns
Dead Time (DT)	0.5		3.33	ns
Readout speed		300		MSample/s

Table 3.4: Performance summary

3.3.4 Discussion

In principle, FPGAs are not designed to implement TDCs but they are particularly well suited to combinatorial and sequential logic. For this reason, the timing margins for the FPGA components are not specifically optimized for uniformity or reduced jitter. While we already discussed timing non-uniformities due to clock skew and routing delays, we describe our approach to jitter reduction. Jitter in our design has a direct impact on the uncertainty of the measurement and comes from two distinct sources. First, while propagating through the delay line, our signal accumulates jitter. In fact, every time it passes through a gate or buffer the timing uncertainty of the transition is affected. The only way to limit this effect is to reduce the number of elements in the delay line. However, as already noted, this implies increasing the clock frequency in order for the period to be covered by the delay line. The second source of jitter is related to the clock and its distribution. The reference clock is generated on board by an 30 ppm crystal oscillator fed into a frequency synthesizer (ICS843001-21) that drive a differential clock buffer (ICS8543) before entering the FPGA through a differential clock input pair. The measured the jitter of the clock period is 11.02 ps. Improved clock jitter timings can be achieved with a custom board design with carefully chosen components [82]. Virtex-5 Digital Clock Managers (DCMs) provide convenient clock facilities of clock multiplication/division and phase shifting. However we refrain from using DCMs because of the potential jitter introduced. For these components, the clock out jitter estimated by Xilinx Architecture Wizard (arwz) is of about 100 ps.

Dummy cells at the beginning of the line are also used in this work, albeit not explicitly presented. The load asymmetry of the first elements cause non-linearities that should be avoided. This is a well known procedure in traditional ASIC TDC design.

The designer of a TDC should take particular care to metastability issues. The signal to be measured will eventually produce a transition that violates setup or hold conditions in the system. The use of synchronizers generally mitigates the effects of metastability. In our design, note that flip-flops that become metastable are limited to those that latch the delay line. In fact the hit signal in figure 3.15 only drives the first filtering flip-flop which in turn drives the delay line. The detection of an event is done in the encoder by comparing two successive states of the latched delay line. In this manner, we prevent metastability to propagate into the system.

Architectural changes could provide enhancements in two directions: linearity and resolution. The pseudorandom bin dithering technique presented in this thesis addresses the linearity issue and we showed that an INL range below 1 LSB can be achieved. The technique however is suited only for measurements of a repetitive events. If single-shot measurement is required the nonlinearity should be controlled. Downsampling as presented in this work comes at the expense of resolution. The INL could be improved by reducing the delay line length as described in section 3.3.4. The expected INL (fig. 3.27) was derived from our current system by using only 64 elements. Reducing the line length requires either to increase the clock frequency or to duplicate the line and work on the both clock edges.

In order to improve resolution one needs to take advantage of the "smaller" bins. The approach taken in [79] is to generate extra transitions with an FIR or IIR filter in order to catch at least one transition in a smaller bin. Another way would be to use several delay lines and control the arrival of delay of the hit signal in each line. As devices with newer technology (Virtex-6: 40 nm, Virtex-7: 22 nm) become widely available, we expect the scaling will be favorable to resolution.

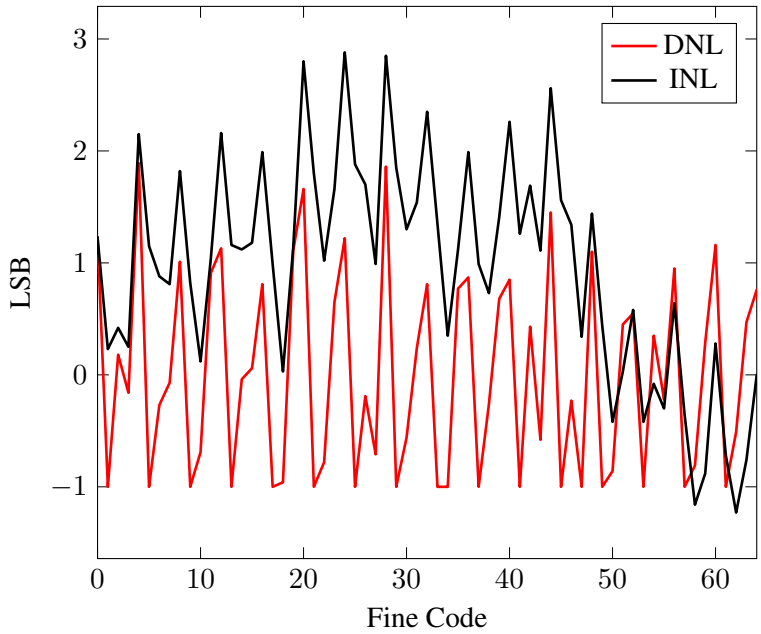


Figure 3.27: Expected DNL and INL (simulated) of 64 bins delay chain. Note that the large nonlinearity at bin 75 would be avoided in this shorter delay line.

3.3.5 130 nm FPGA TDC implementation

The main differences between the 65 nm Virtex 5 and 130 nm Virtex II implementations are summed up in table 3.5. Besides the obvious technology change, the 130 nm implementation uses 1 tap per slice compared to 4 taps per slice in the 65 nm TDC. The resolution change is mainly due to this implementation detail. The nonlinearty of the Virtex II implementation is reported in figure 3.28. Note that the large DNL variations are also present.

	Virtex II	Virtex V	Unit
Technology	130	65	nm
Resolution	57.6	17	ps
Taps/Slice	1	4	
Clock Freq.	200	300	MHz
DNL Range	$[-0.8, +0.5]$	$[-1, +3.55]$	LSB
INL Range	$[-0.45, +1.47]$	$[-2.99, +2.58]$	LSB

Table 3.5: Comparison of Virtex II and Virtex V implementations.

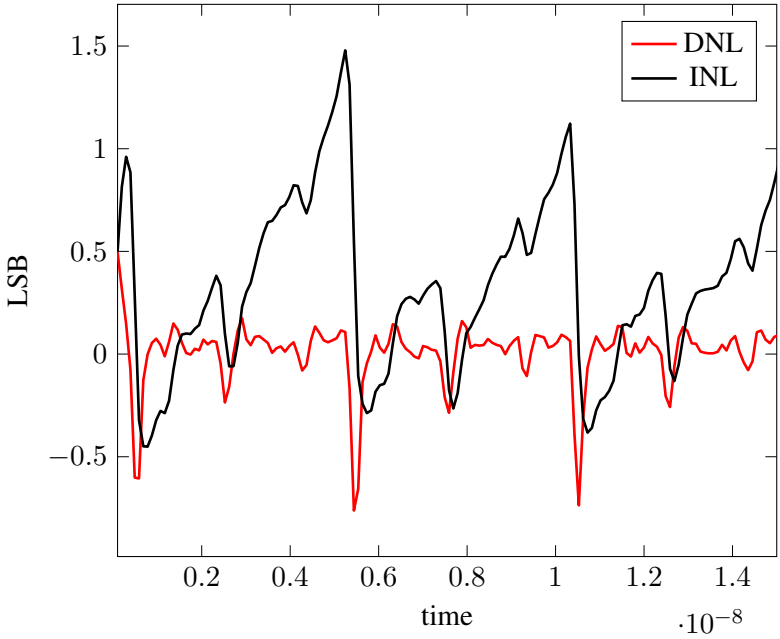


Figure 3.28: DNL and INL of the TDC architecture implemented in a Virtex II device.

3.4 Conclusions

A method for inter-/intra-chip optical communication was presented in this chapter. The proposed modulation, Pulse Position Modulation, was selected with the SPAD dead time and noise, in mind. Theoretical limits on the achievable bandwidth were derived in section 3.2. A detailed implementation of time-to-digital converters in FPGA was presented as a possible building block of the demodulation part. As demand for short range high bandwidth communication grows, we believe that highly integrated optical solutions will be pursued in the future as the research progresses in this field [2,3].

4 SINGLE-PHOTON PROCESSING AND READOUT

PHOTON detection alone is meaningless if not combined with proper ways to transfer and process the flux of impulses generated. The context of processing is that of defining what information the photons are carrying. There are three processing techniques that we will discuss in the following sections: time-uncorrelated, time-correlated, and spatio-temporal correlated processes. The main aspects will be described along with some known applications.

Readout refers to all the methods of organizing and transporting either the photon induced pulses or the information derived thereof. Processing techniques and readout methods are not independent. In fact, given a processing technique, a selection of readout methods will be more appropriate than another. However, for the sake of clarity, we will first present both techniques separately and then illustrate their interactions with a few cases studies in section 4.3.

Most of the examples in this chapter are related to imaging. However the processing and readout techniques presented are not limited to this application. As a matter of fact, the single-photon communication paradigm introduced in chapter 3 can be seen as a time-correlated technique.

4.1 Processing Techniques

4.1.1 Time-Uncorrelated Techniques

Counting photons is the first and simplest example of a time-uncorrelated technique also called Time-Uncorrelated Photon Counting (TUPC). In this case, photon triggered pulses are used to incre-

ment a digital counter. Decrementing counters is also possible, in some cases, when photon count differences are needed. This accumulation can be compared to the traditional charge-coupled devices (CCD) or CMOS image sensors. The intensity of the optical signal can be retrieved after a period of time – the equivalent of the exposure in photography – with extremely high dynamic range (DR). In fact the DR is only limited by the counter width, readout, speed, and dead time. For example, the design in [103] uses a 1 bit counter in each pixel and reaches a dynamic range of 90 dB when operated at high speed.

Time-gated acquisition is a photon counting technique which uses N counters enabled during orthogonal time-windows of width Δ_t as shown in figure 4.1. The range $N\Delta_t$ is usually chosen to match the repetition rate of the illumination system. The value of the counters after sufficient photon accumulation approaches the optical waveform desired.

Single-Photon Synchronous Detection (SPSD) is a time-gated technique, however, in this case, the system requires that the illumination have a well known temporal shape (a sine curve, for example). By using a time-gated technique, the mean intensity, mean amplitude, and phase shift can be derived. A system implementing SPSD will be described in more detail in section 4.3.2.

4.1.2 Time-Correlated Techniques

Time-correlated techniques make use of a time reference, either optical or electrical to which a measurement of time is referred. The time reference is usually a pulsed laser used to illuminate a sample. We present here two time-correlated techniques, TCSPC and time-of-flight.

TCSPC is a technique that allows the retrieval of an optical waveform by "accumulation" of the time-of-arrival (TOA) of pho-

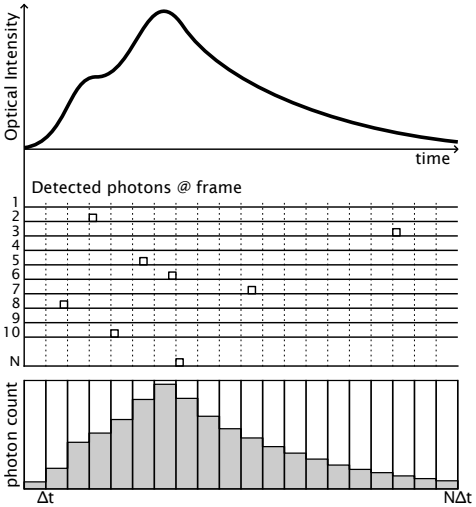


Figure 4.1: Time-gated and Time-correlated single-photon counting principle of operation. The optical waveform emission has to be triggered many times in order to have a meaningful statistic. This is usually done with a short pulsed laser diode (a few picoseconds or less) at repetitions rates between 1 MHz and 100 MHz. For time-gated N counters each covering a Δ_t window are incremented when a pulse lies in the corresponding window. The value of all the counters next to each other represents the waveform. For Time-Correlated Single-Photon Counting (TCSPC), Δ_t is the resolution of a TDC that measure the time of arrival of the pulses generated by a pixel. The histogram of all TOA values represents the desired waveform.

tons [104, 105]. It assumes that the desired waveform can be acquired several times. A repetitive reference stimulus such as a short pulsed laser beam (a few picoseconds or less) is usually employed to trigger the waveform to be analyzed at repetitions rates between 1 MHz and 100 MHz. The time of arrival of photons with respect to the reference is used to build an histogram. This histogram rep-

resents the statistics of the photon arrival or in other words a *scaled* image of the variation in time of the intensity of the optical waveform. The principle of operation of TCSPC is shown in figure 4.1.

The Time-of-Flight (TOF) technique, like TCSPC, measures the time-of-arrival of a photon with respect to a fixed reference. The measurement, however, is meant to precisely retrieve the distance of an object or scene. The pulsed illumination required can be a laser and the system measures the time-of-flight of photons that are reflected back from the object. A complete description of a system designed for TOF 3D imaging is presented in the case study section 4.3.1.

4.1.3 Spatio-Temporal Correlated Techniques

In spatio-temporal correlated techniques the signals $s_1(t)$, $s_2(t)$, $\ldots$, $s_N(t)$ over a measurement period T_{meas} are recorded from N sensors. Let $I_i(t)$ be the number of photons recorded in signal $s_i(t)$ in a window of time Δ_t centered around time t. The auto correlation $g_{ii}^{(1)}$ is defined by

$$g_{ii}^{(1)}(\tau) = \frac{\langle I_i(t) I_i(t+\tau) \rangle}{\langle I_i(t) \rangle^2},$$

where $\langle . \rangle$ is the average over T_{meas}. The n^{th}-order spatio-temporal cross-correlation function $g_{ij}^{(n)}$ is

$$g_{ij}^{(n)}(\tau) = \frac{\langle I_i(t) I_j(t+\tau)^{n-1} \rangle}{\langle I_i(t) \rangle \langle I_j(t)^{n-1} \rangle},$$

Depending on the application, the value of $g_{ij}^{(n)}$ yields some specific information on the system. In Fluorescence Correlation Spectroscopy (FCS) [106], the auto-correlation ($g_{ii}^{(1)}$) yields information

on molecular concentration, motion, state, etc. Higher-order techniques have also been used like in [107] but using a single measurement signal. The second order spatio-temporal correlation $g_{ij}^{(2)}$ might be used to confirm the presence of true Bose-Einstein macroscopic coherence (BEC) of cavity exciton polaritons [108, 109]. In this case, the signals from two or more SPAD pairs are analyzed.

4.2 Readout Strategies

Dense arrays of SPADs with pitch as low as a few tens of micrometers have been fabricated. Thanks to the large amplification gain in Geiger mode, the in-pixel circuitry can be minimal compared to the large area-consuming trans-impedance amplifiers necessary with other sensors. Photon detection probabilities as high as 40 % allow practical applications of TUPC, TCSPC, and spatio-temporal correlated techniques.

Besides the evident photography and video recording applications, SPAD based imagers open the way to applications where large processing power is highly integrated with the detection process. The integration of TDCs (also see section 3.3) is a good example of integrated processing used in TCSPC applications such as 3D ranging cameras and Fluorescence Lifetime Imaging (FLIM).

Readout becomes particularly important when large and dense arrays of single-photon sensors are built. While the size is pursued to achieve high resolutions, density is required for fill factor efficiency. The fill factor is the ratio between total active area and total sensor area. For example, guard ring structures, quenching, recharging, and buffering circuitry are in-pixel components that contribute to fill factor losses. Note also that large active area pixels are not necessarily preferred. In fact, large area pixels generally exhibit larger jitter and noise figures in CMOS.

The readout strategy also impacts the acquisition time and therefore the frame rate. For a given application, it is generally required that N measurements or photon arrivals per pixels are collected. The acquisition time τ_{acq} and illumination are set such that this condition is met. If the readout strategy imposes that only a part of the array is active at a time, the acquisition time will increase. Readout time and acquisition time are usually interleaved or parallelized when possible in order to maximize throughput.

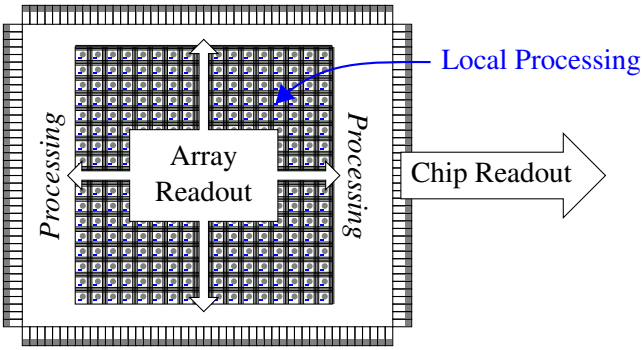


Figure 4.2: Imager chip illustration, the sensor array at the center can contain local processing elements. The pad ring provides the interface to the outside for chip readout.

We distinguish the readout process between two parts: from the pixels to the boundary of the array (array readout) and from the array's boundary to outside of the chip (chip readout). The main difference between the two is that in the first case the area used by the readout circuitry impacts the fill factor while this is not the case in the second. Figure 4.2 illustrates the typical layout of sensor chip, array readout, processing, and chip readout.

4.2.1 Array Readout

The two extreme readout schemes are the fully parallel strategy and the fully sequential strategy. In the former, every pixel has a dedicated routing resource to the edge of the array hence frame acquisition time and readout is minimized at the cost of a large routing area scaling approximatively with $O(N^{0.5})$ where N is the number of pixels. In the latter, the information from only a single pixel at a time can be read. For example the random access readout scheme allows selection and readout of a single pixel addressed by its coordinates. This rather flexible scheme uses silicon area sparingly at the cost of large readout time due to sequential access. This strategy yields the slowest frame rate and doesn't take advantage of any parallelism or resource sharing. In fact, by introducing parallelism and resource sharing, several strategies can be devised to bridge between these two extremes.

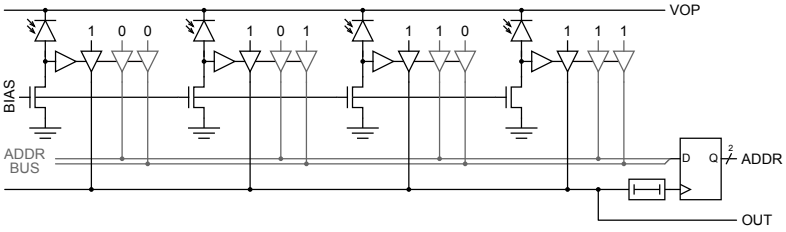


Figure 4.3: The event-driven readout in-pixel circuitry requires $1 + log_2(K)$ transistors, where K is the number of shared pixels. The tristate buffers that drive the externally pulled-up address bus are generally able to only pull the lines low (only 1 transistor required). [110]

In the event-driven scheme introduced by Niclass *et al.* in [7, 110, 111], K pixels share a common readout line and a digital address bus as shown in figure 4.3. The event-driven readout in-pixel circuitry requires $1 + \log_2(K)$ transistors, where K is the number

of shared pixels. In order to minimize the number of transistors, open-drain signaling is used and the address bus and readout line are pulled-up with a resistor at the periphery of the array or externally. This scheme assumes that photon detection and pixel noise are low in order to reduce the probability of collisions on the readout bus. In fact, a collision on the address bus may lead to latching wrong values when the setup or hold time of the flipflops on the address bus are violated. Let $\Delta T = t_{\text{setup}} + t_{\text{hold}}$ be the sum of the setup and hold times of theses flipflops. Given the light intensity Φ_{s} and the photon detection probability p_{pd}, the probability of a collision in the window ΔT can be computed as

$$P(\text{Collision}) = 1 - \epsilon_{\text{pdp}}{}^{K-1},$$

where $\epsilon_{\text{pdp}} = e^{-\Phi_{\text{s}}\Delta T p_{\text{pd}}}$, as defined in chapter 3, equation 3.11. $P(\text{Collision})$ is plotted in figure 4.4 as a function of K. The value $\Delta T = 100\,\text{ps}$ is a typical setup and hold time window of a flipflop for a 0.35 µm CMOS process.

This calculation assumes uniform and uncorrelated light over the K shared pixels, which might not always be the case. As a first approximation, background noise and SPAD DCR can be considered to artificially increase the light intensity Φ_{s}.

While the collision probability is kept to a minimum by controlling the light intensity, collisions could be detected with the addition of some logic and routing. For example, pixel addresses could be coded on more bits and chosen such that a collision would lead to an invalid address. For example, 6 pixel addresses could be coded on 4 bits with the following alphabet: $\{1100,\ 0110,\ 0011,\ 1010,\ 0101,\ 1001\}$. This technique resolves up to two colliding photon detections. The number of additional bits in the alphabet is related to the maximum number of colliding photon detections one wants to resolve. Another way of detecting collisions would be to have a

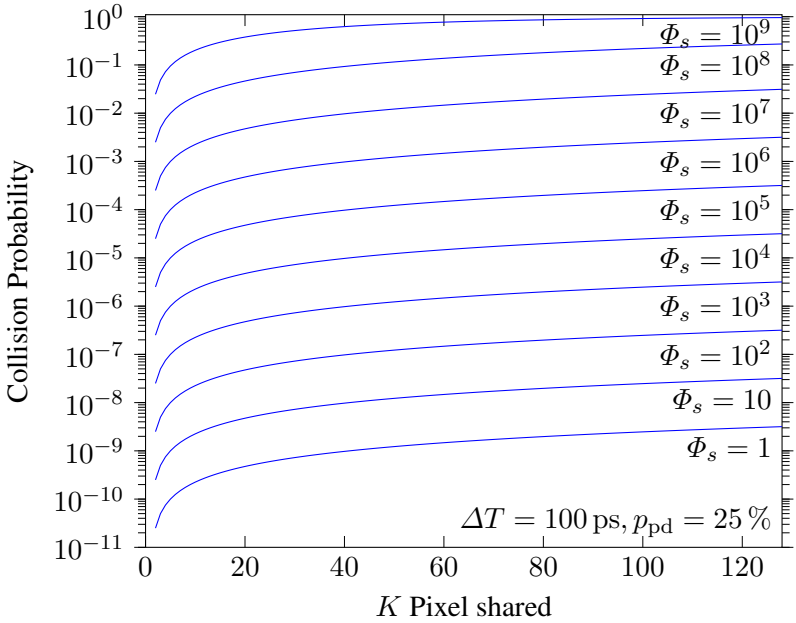


Figure 4.4: Collision probability as a function of the number of shared pixels. $\Delta T = 100\,\mathrm{ps}, p_{\mathrm{pd}} = 25\,\%$.

1-bit latch in the pixel that is set when the SPAD triggers, and build a shift register out of all shared pixel in order to check that only 1 pixel fired. This technique allows to detect and discard any number of collisions but it requires more time to read out the complete shift register image.

The main advantages of the event-driven technique are that the acquisition time per pixel can be minimized, readout occurs opportunistically, the number of line required for readout grows with $O(\log_2(N))$, and a single measurement device such as a TDC can

also be used for all the shared pixels*. Note that in this last case, the dead time of the TDC limits the measurement rate (some photon generated pulses are lost).

The average pulse rate Ω is a measure of how many information carrying pulses are generated on average per unit of time. In the case of K ideal SPADs,

$$\Omega_{\text{ideal}} = K \cdot \Phi_{\text{s}}.$$

For the event-driven readout, collisions limit the pulse rate proportionally to the collision probability derived earlier.

$$\begin{aligned}\Omega_{\text{ed}} &= K \cdot \Phi_{\text{s}} \cdot P(\text{No Collisions}) \\ &= K \cdot \Phi_{\text{s}} \cdot e^{-(K-1)\Phi_{\text{s}}\Delta T p_{\text{pd}}}.\end{aligned}$$

The dead time τ_{dt} of the SPAD also impacts the pulse rate. The average number of pulses produced in a windows of time τ_{dt} is $1 - P(\text{No pulses generated})$ hence the average pulse rate of K SPADs:

$$\Omega_{\text{dt}} = \frac{K(1 - e^{-\Phi_{\text{s}}\tau_{\text{dt}}p_{\text{pd}}})}{\tau_{\text{dt}}}.$$

When a TDC is shared amongst K pixels, the TDC dead time τ_{tdc} is a limiting factor for the average measurement rate Ω_{tdc}. Assuming that no pulses are lost due to collisions such as those in the event-driven readout and that the SPAD dead time is ideal, the average measurement rate Ω_{tdc} is

$$\Omega_{\text{tdc}} = \frac{1 - e^{-K\Phi_{\text{s}}\tau_{\text{tdc}}p_{\text{pd}}}}{\tau_{\text{tdc}}}.$$

In figure 4.5, the average pulse rate was normalized to the number of shared pixels K. The first limiting factor is the TDC dead

*A shared TDC also provides uniformity in measurements across the pixels.

time and is reached around 1×10^6 photons/s in this example. At higher illumination, the event-driven collision rate is bringing down the pulse rate. The SPAD dead time becomes a limiting factor when no event-driven readout is employed or when the TDC dead time is lower than the SPAD's dead time. Note that for clarity in the figure the normalized Ω_{tdc} is shown only for $K = 2$.

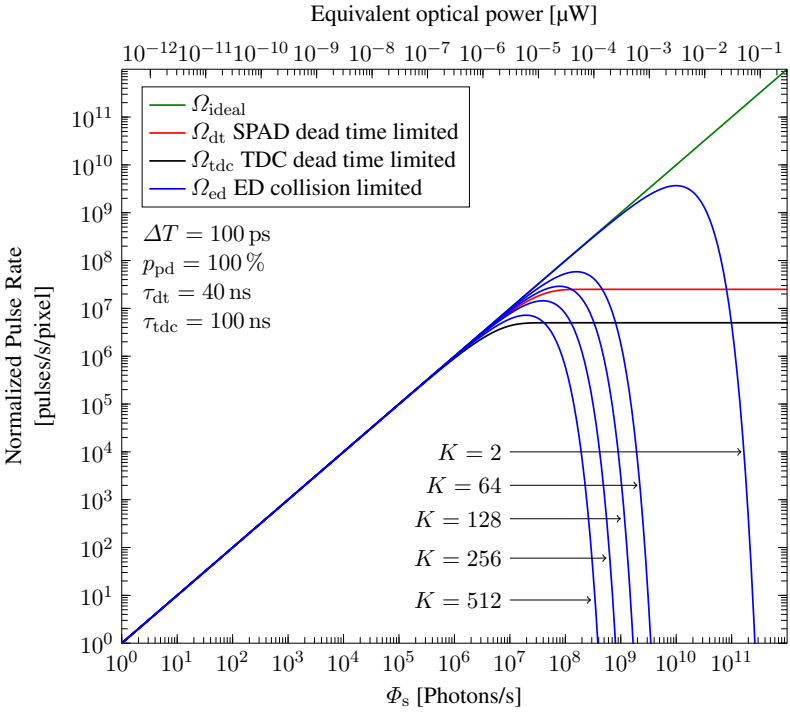


Figure 4.5: Normalized pulse rates $\Omega_{\mathrm{ideal}}, \Omega_{\mathrm{dt}}, \Omega_{\mathrm{tdc}}, \Omega_{\mathrm{ed}}$ as function of the illumination Φ_{s}. Note that Ω_{tdc} is only shown for $K = 2$. The equivalent optical power axis was calculated for a wavelength of 650 nm. Parameters: $\Delta T = 100\,\mathrm{ps}, p_{\mathrm{pd}} = 100\,\%, \tau_{\mathrm{dt}} = 40\,\mathrm{ns}, \tau_{\mathrm{tdc}} = 100\,\mathrm{ns}$.

In the row-based rolling shutter readout technique, a single line of pixels is read out at a time. After a defined period of time τ_{l}, the successive line is selected. Pixels of the same line share the same activation signal that controls a transmission gate that drives a column shared readout line. If an acquisition time τ_{acq} is required per pixel, $\tau_{\mathrm{acq}}/\tau_{\mathrm{l}}$ cycles will be required for a complete frame. Between row readouts the pixels can continue accumulating counts if a local counter is present on-pixel [103]. Column based rolling shutter is, of course, also possible.

The region-of-interest (ROI) readout is a particular case of readout that can be applied to any of the schemes above. In order to increase the frame rate, a subset of lines and columns can be selected thus reducing the amount of data to be transferred. For example, this technique is implemented in the design presented in [112].

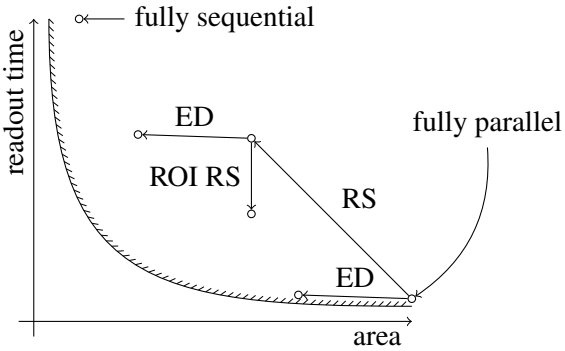


Figure 4.6: Tradeoff between time of acquisition and readout of a complete frame (or ROI) and the area of the readout logic. Starting from a completely parallel design, note the effects of event-driven (ED) and rolling shutter (RS) techniques.

When we look at the tradeoffs between acquisition and readout time of a complete frame and the area required for the readout

logic, picture 4.6 gives an overview of the advantages of each strategy presented. Area savings in the array can potentially be obtained with the event-driven scheme. However one the major benefits of this readout strategy is that it allows sharing of resources such as TDCs in TCSPC applications. The same is applicable to the rolling shutter strategy. Note that in ROI readout, the time is reduced because of fewer pixels are read out.

4.2.2 Chip Readout

The size of the periphery of the array is generally *only* constrained by area and power budgets. Processing of the sensor signals may be performed in this area with TDCs, memories, or even processors. Configuration logic, biasing circuitry, clocking, and chip readout circuitry are generally also present in this region.

Access to the outside of the chip is done through the pads which are usually wirebonded as flip-chip bonding is not practical for such sensor chips. Whether the design is pad limited or core limited, the chip readout may be composed of multiplexing, buffering, compressor, and serializer logic elements. For most prototypes, the processing is moved outside the chip to a FPGA where large memories and many logic elements are available conveniently and at low cost.

4.3 Case studies

The two following case studies (and their respective ICs) illustrate how processing techniques and readout strategies can be combined. The readout systems of both chips were implemented by us in the framework of a collaborative project that yielded the ICs. Both examples are implementations of 3D ranging cameras. A 3D ranging camera is a device that measures the distance from the sensor's pix-

els to a specified object or scene. Applications of 3D ranging cameras include human-computer interfaces, robotics, biometrics, and automotive to name a few. According to the classification in [111], solid-state 3D image sensors used in TOF cameras can be categorized by the optical modulation used to illuminate a scene: pulsed modulation and continuous modulated wave. Pulsed modulation is shown in the first case study, while continuous modulated wave illumination is shown in the second.

4.3.1 Case study 1: LASP 3D camera

The basic principle of pulsed optical modulation is to measure the *time-of-flight* of photons generated from a light source and that are reflected by the scene back to the sensor.

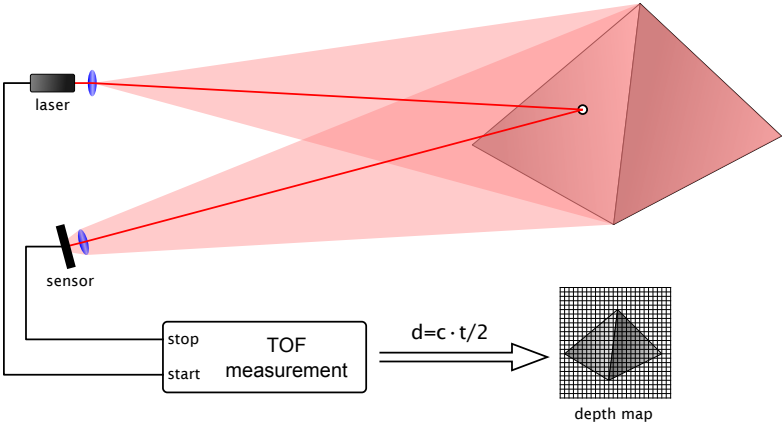


Figure 4.7: Principle of operation of a pulsed modulation 3D camera. The laser emits short pulses of light which are reflected back to the sensor. The time between the laser trigger pulse and the photon detection at each pixel of the imager is used to resolve the distance.

Figure 4.7 pictures the setup of a pulsed modulation TOF camera. The laser in this example is pulsed at a fixed frequency while the imager senses the arrival of the photons reflected from the scene. For each pixel of the sensor, a measurement t_{tof} of the time between the laser pulse and the peak of the photons arrival at the sensor is taken. The distance d for each pixel is then calculated by

$$d = \frac{c \cdot t_{\mathrm{tof}}}{2}$$

where c is the speed of light. Several measurements are gathered in histograms or averaged to refine the confidence interval before a frame is transferred for display or further processing on a commodity computer or on an embedded display.

Prototypes of pulsed modulation ranging cameras have been demonstrated [111, 113–115]. Accuracy of the measurements in the order of 1 mm to 2 mm is achieved by averaging 10^4 samples in [114]. A large array of 128×128 pixels with 25 µm pitch and 32 integrated TDCs was built for this purpose [115]. While the precision is of 9 mm and accuracy 5.2 mm mainly due to the internal TDC non-linearity, the great advantage of this imager is it's very short acquisition time: 50 ms per frame.

The architecture of the 128×128 array of SPADs with integrated TDCs reported in [115] is presented here. The scope of this case study is to discuss the readout architectural decisions made in this design by us. The original goals were to provide a relatively compact imaging platform for time-correlated and time-uncorrelated photons counting measurements with timing precision in excess of 100 ps and frame rates as high as 1 kHz.

Ideally time measurement would be done in parallel for every pixel. This would imply that the TDCs would occupy a very large area and the benefits of parallelism would be tainted by a possible non-uniformity in measurement. The choice was made to limit the

number of TDCs at the expense of a limited array readout speed. This limitation impacts the readout in two ways. First, event-driven readout as presented in [7, 110, 111] is used. The number of transistors per pixel to implement this technique is $\lceil 3 + \log_2 N \rceil$ where N is the number of shared pixel. These transistors are added to the 7 transistors required for the active quenching, active recharge, and buffering. Also the more pixels shared, the more addressing lines are required which require space in the floorplan. In the end, $N = 4$ pixels across a single row were chosen. Second, a rolling-shutter row selection was implemented to organize the array readout. Note that, due to the dynamical nature of SPADs and of the rangefinding process, only the currently selected pixels could be active. The same is not true in designs where the counting of photons is performed in each pixel independently and can be held in a local counter or memory as in [103]. It is important to understand that these choices were done iteratively taking area (hence cost), and performance in account. The choice of limiting the number of TDCs directly impacts the rate at which each pixel light sensing activity is used.

Limiting the number of TDCs also reduces the required readout bandwidth. As a rough calculation, the bandwidth required if there were 128×128 TDCs with 10 bits resolution at 20 MSamples/s would be of 3276.8 Gbps. The first approach taken to reduce this figure was to compress and/or build histograms on chip. The memory requirements for histogram building were considered. When accounting 1 kB of memory per pixel, the required area for 128 kB dual ported memories on the targeted 4 metal layers 0.35 µm high-voltage CMOS technology was approximately 60 mm^2. This number compared to 10 mm^2 of the pixel array pushed for a more compact solution where efficient use of the silicon would be pursued. Therefore, solutions to put the histogram generation outside of the chip required addressing the readout bandwidth problem.

The readout of 32 TDCs requires a bandwidth of 7.68 Gbps. For design safety margin and power consumption, it was considered that a single-ended pad could sustain 80 MHz operation of a line capacitance of 10 pF (input capacitance of a FPGA pad) with 4 mA drive strength. Therefore 96 pads would suffice for this design. The TDC readout is time-multiplexed in order to provide the data of 8 TDCs sequentially. Figure 4.8 shows the TDC to pad readout interface. Note that, to save power, output pad data is only changed when valid.

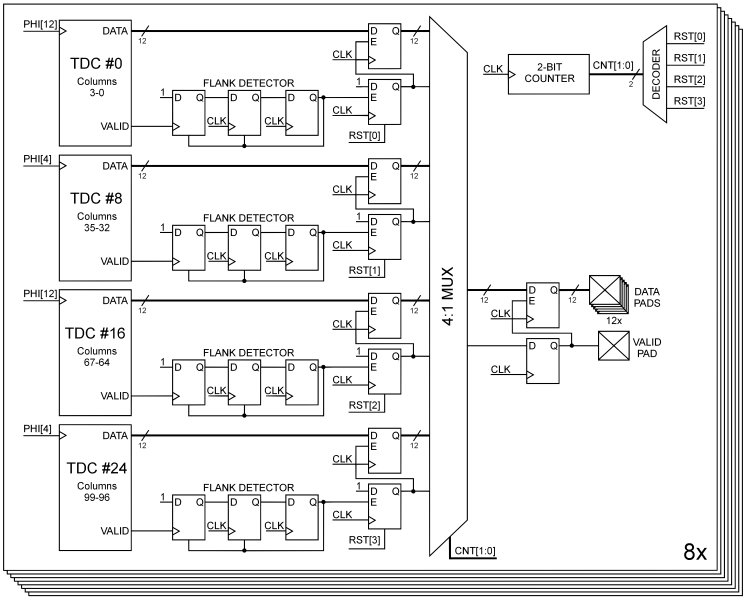


Figure 4.8: Chip readout synchronization and pad interface. The readout clock (clk) run 4 times faster than the TDCs. Note that in order to save power, the data pads value is changed only when valid. Drawing credits: C. Niclass.

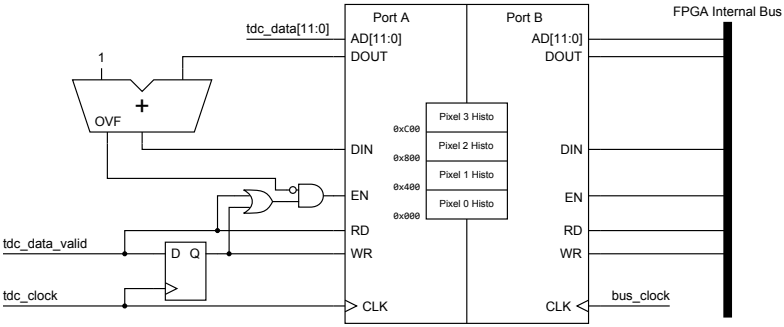


Figure 4.9: Simplified schematic of an histogram builder. A dual ported memory is used to adapt for frequency domains.

The flexibility of letting an FPGA post-process the raw data allows us to use a more advanced technology (130 nm CMOS for the Virtex II device used), increases yield, and lowers cost. Histogram building in FPGA is conceptually trivial, as large dual-ported memories are available. A simplified schematics of histogram building circuitry is shown in figure 4.9. Further data transfer to a computer were possible through a variety of protocols such as USB or Ethernet. A dedicated motherboard was designed with Virtex II devices (see figure 4.10).

The complete system is shown in figure 4.11. A boundary scan controller (JTAG) is also available to perform debugging of the readout without compromising normal running operations. The chip shown in figure 4.10 was used for range finding applications [115], microlense testing, and it is in currently studied for use in a prototype optical near infrared brain scanner.

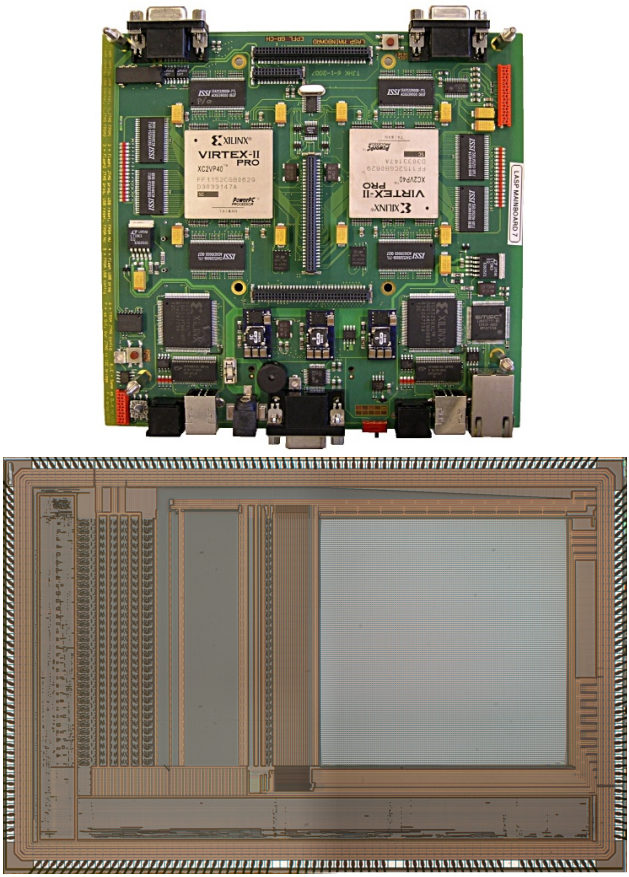


Figure 4.10: Top: The dual FPGA board designed for histogram processing and data transfer to a computer. Bottom: Array of 128×128 SPADs fabricated in 0.35 µm high-voltage CMOS technology. The chip measures $8 \times 5\,\mathrm{mm}^2$ with a pixel pitch of 25 µm. [115]

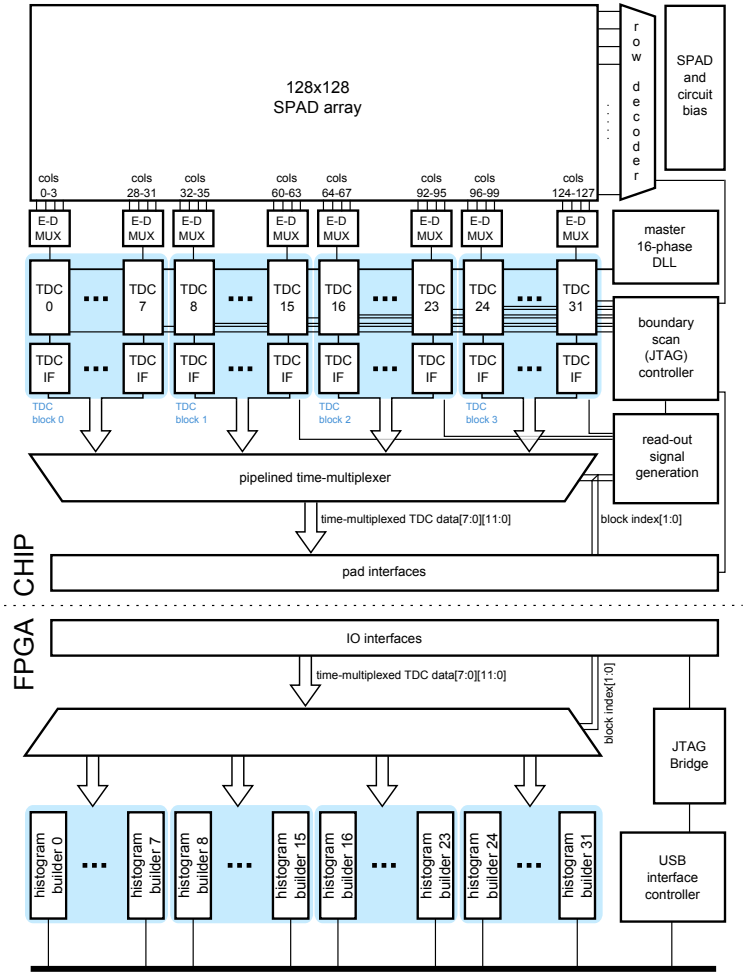


Figure 4.11: System architecture of the 128 × 128 array of SPADs with integrated TDCs. The readout is interfaced with an FPGA for histogram building and post-processing. Transfers of data to the computer are performed through USB 2.0.

4.3.2 Case study 2: SPSD 3D camera

The principle of operation of a continuous modulated wave TOF camera is to measure the phase shift induced by the time-of-flight of photons generated by a reference frequency f intensity modulated light source. Per pixel, the receiver uses homodyne detection principles where the reference source and its quadrature are used to resolve phase and amplitude. From the phase φ, the depth d is calculated as follows:

$$d = \frac{c}{2} \cdot \frac{\varphi}{2\pi f}.$$

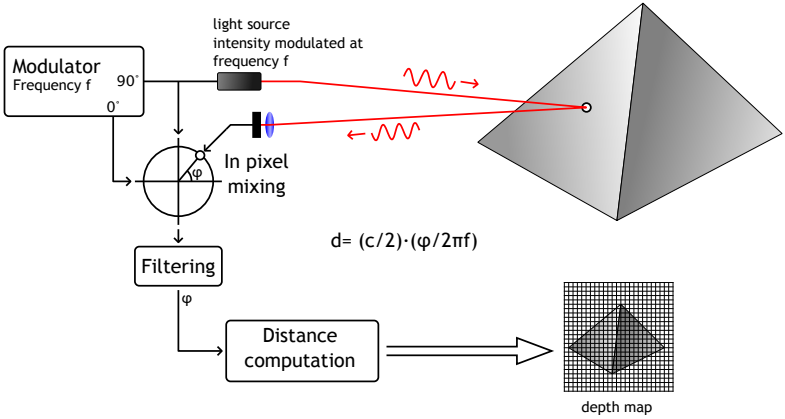


Figure 4.12: Principle of operation of a continuous wave time-of-flight measurement system. The phase shift of the received optical signal is measured with *electrical homodyne detection*.

An overview of the system is shown in figure 4.12. A continuous modulated wave SPAD-based imager was shown in [116] and [111]. In this work, a frame is captured in 45 ms[†] and the

[†]Recently this figure was decreased to as low as 10 ms

illumination system is composed of 48 infrared LEDs producing 800 mW of optical power. Practically, the system composed of 60 by 48 pixels uses two 8-bit counters per pixel yielding a pitch of 85 µm.

The readout strategy uses rolling row shutter and 8-columns blocks parallel multiplexing. The data from the chip is transfered to an FPGA for further processing and control of upload to a PC through an USB interface.

The camera achieves a worst case precision of 11 cm and worst case accuracy of 3.5 cm. This seemingly low performance is partly due to the illumination setup, moreover systematic errors have not been removed by calibration techniques. Therefore, a large performance improvement is to be expected in the future with this technique.

4.4 Conclusion

In this chapter, we presented single-photon processing techniques and readout strategies. Three processing techniques were detailed: time-uncorrelated, time-correlated, and spatio-temporal correlated processes. The readout strategies presented covered readout of the array and readout of the chip. The tradeoffs of the different readout schemes were discussed and analyzed. The case studies presented illustrate some applications of these principle that were implemented in silicon.

5 SINGLE-PHOTON CLOCKING

5.1 Motivation

OPTICAL clock distribution has been a subject of active research for the past two decades. Even as early as the 1980s, with the rise of fiber-optics in telecommunications, Goodman *et al.* were the first to present a thorough analysis of optical interconnects for VLSI systems [117]. However, conventional electrical distribution remains the norm to date. The reasons for this trend were that very fast detectors in standard CMOS processes were not available until recently. In addition, the need of *ad hoc* packages for optical distribution and fiber coupling deterred most manufacturers to pursue the optical clock route for cost and compatibility reasons.

Optical means for clock distribution and data transfer directly on chip are attractive for a number of reasons. In *primis*, an optically coupled network is less subject to the usual performance limitations of its electrical counterparts, such as skew, jitter, and power consumption, especially at high frequencies. In addition, with the emergence of 3D integration, fast through-chip communication and clocking has become a real issue, whereas through silicon vias, the currently proposed solution, are too bulky as of 2010 although they seem to start being effectively and reliably mass-produced. On the contrary, an optical channel can be implemented today through a stack of thinned silicon chips using conventional micro-optics techniques and air or dielectric based waveguides (figure 5.1). Silicon dioxide and germanium waveguides can be used in a planar chip for horizontally pushing optical pulses. Their fabrication is becoming

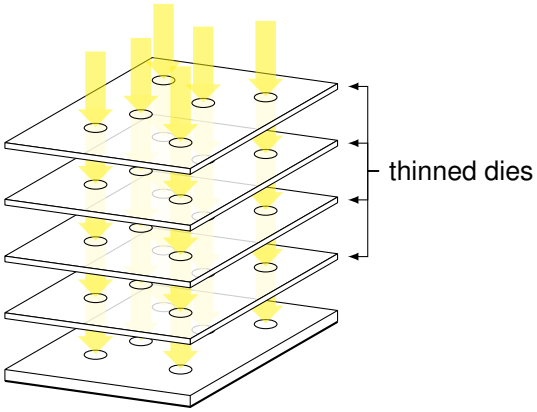


Figure 5.1: Optical channels in a stack of thinned chip.

common-place and it is already CMOS compatible at least in some SOI technologies [37].

Optical clock distribution can provide reduced skew and jitter in distributing the synchronization signal, though not necessarily slashing power [118]. A remaining problem is that of optical-to-electrical conversion [119] which [118], [120], and [121] have tried to solve with some success. However, much remains to be done in this field even with the encouraging advances achieved in optical clock distribution at the chip, package, board, and cabinet level [117, 122–125]. Interesting new directions are currently being pursued using alternative waveguide materials, such as germanium that can be used horizontally and vertically. Fabrication of these waveguides can be performed even at low temperatures, thus making them compatible with a post-processing step on advanced deep-submicron CMOS technologies.

The development of purely electrical, high-performance clock networks has meanwhile progressed in the last years, yielding sche-

mes to locally generate high-frequency clock signals in the spirit of a Globally Asynchronous Locally Synchronous (GALS) approach. The solution generally adopted is that of a closed-loop-with-active-compensation that is implemented by means of Phase-Locked Loops (PLLs) and Delay-Locked Loops (DLLs) [126]. In this context, much effort has been devoted to reducing jitter and power [119, 127–130]. However, these techniques are also generally power- and area-hungry. Besides PLL and DLL based circuits, ring oscillators have been proposed. With their simple design, compactness and predictable performance behavior [131], these circuits are commonly used for localized clock (re)generation [132–134].

We propose a CMOS optical clock distribution scheme named *Oscar*. Its application, detailed in section 5.3, is adapted to, but not limited to synchronization of an embedded processor with Instruction Set Extensions (ISEs) implemented on chip. The particular Custom Instruction Units (CIUs) proposed in this chapter are circuitries that perform logic and arithmetic operations at an internal clock frequency that is significantly higher than that of the processor they serve. The system can be thought of as fully synchronous but without the burden of global high-speed clocks that are replaced by ultra-low-power optical clock pick-ups based on single-photon detectors (fig. 5.2). The single-photon detectors used in this work are CMOS compatible SPADs that were developed for the first time in a sub-100 nm CMOS technology by our group [14]. At the time of the design, the SPAD ensemble was choosen at the same time that the first results of the 90 nm SPADs were available. We discovered only later that the selected SPAD ensemble is not functional. However, for the purpose of the following discussion and without loss of generality, we used an external 0.35 µm SPAD.

The use of SPADs for the optical pickups, instead of conventional photodiodes, has several advantages. First, due to the mechanism of self-amplification of SPADs, no amplifiers nor comparators

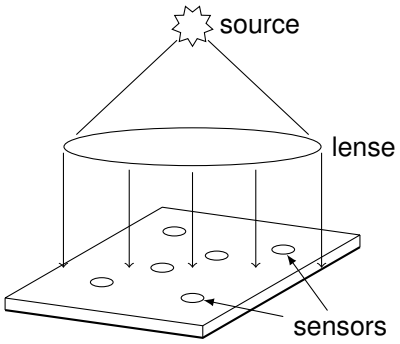


Figure 5.2: Optical clock distribution example with a cone of light over the chip.

are needed to convert optical onto electrical power. In addition, the avalanche process is very fast, thus enabling picosecond resolution in the synchronization edges. Second, thanks to SPAD sensitivity, it is possible to reduce the optical power used at the source and to use a combination of several parallel signals operating in close proximity. Finally, thanks to the miniaturization levels achieved in deep-submicron SPADs, the real estate overhead is negligible [5, 13, 14, 86].

As an alternative optical pickup technology, APDs could be used, the main advantage being an almost inexistent dead time that could enable operating frequencies in the gigahertz range at the price of a relatively complex amplification scheme and very strict bias control circuitry. However, in *Oscar*, global synchronization speeds are not critical, even nanosecond-long dead times are acceptable, as long as the timing resolution remains high, i.e. 100 ps or less. As presented in chapter 2, spurious firing (dark counts) and afterpulsing may occur in SPADs. However, these effects are inherently canceled by *Oscar* architecture.

The chapter is organized in the following manner. First the architecture of *Oscar* is described in section 5.2. Section 5.3 details a possible application that we implemented in our demonstrator. Section 5.4 holds validation considerations (5.4.1), methodology (5.4.2), and results (5.4.3). Finally the discussion in sections 5.5 and 5.6 covers both measurements and future work. The contents of this chapter are adapted from the papers [135, 136].

5.2 Architecture

In this section, we describe the architecture of two implementations of the proposed clocking scheme. The frequency of first implementation is fixed, while it can be selected in a wide range for the second. Both share the same principle of operation which is detailed here after and in figure 5.3. A local oscillator is started by a pulse from the sensor and it is then stopped after an integer number RV of cycles. This mechanism ensures that the edges of all the generated clocks on the chip are aligned at these synchronization time-points. On the other hand, the clock edges in between might not be aligned due to PVT variations. The limitation imposed by this clocking mechanism is that data can only be safely exchanged at the synchronization points.

The constraints in designing the optically synchronized ring oscillator were twofold. First, it should be relatively small and simple so that the area covered is minimal compared to an equivalent "electrical" clock distribution mechanism. Second, it should generate exactly RV rising clock edges without glitches at various frequencies including frequencies above 1 GHz.

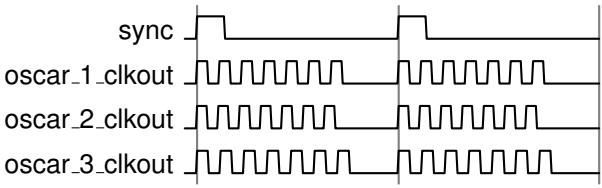


Figure 5.3: Principle of operation of the non-PVT compensated oscillator. The 3 Oscar oscillators situated on different locations on chip are all started at the same time and run for $RV = 7$ cycles. Communication across clock domains is guaranteed only on the synchronization points.

5.2.1 Fixed Frequency *Oscar*

The simplified schematic of *Oscar* is shown in figure 5.4. The output of the D-flip-flop FF1 is used to start the oscillator composed of the NAND-gate and the inverters when a rising edge appears on the SYNC input. The 7-bit down-counter starts from RV (the Reload Value of the counter) after reset and decrements based on a delayed version of the CLKOUT signal. When the counter underflows, the most significant bit is used to reset the D-flip-flop which in turn asynchronously loads the counter with the programmable RV value. Note also that the counter is active only when the oscillator is enabled.

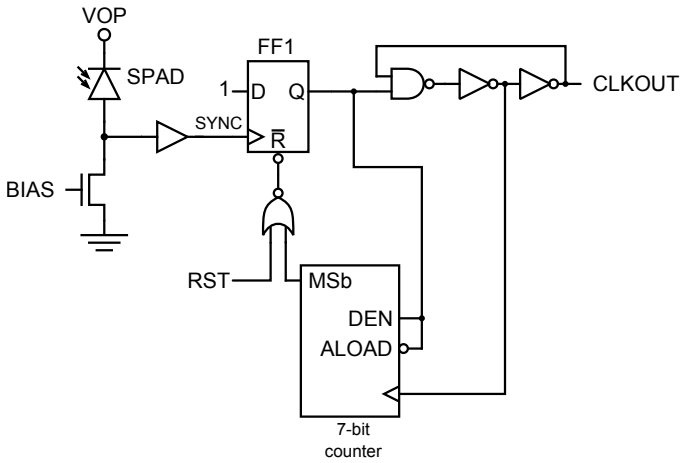


Figure 5.4: Simplified schematic of *Oscar*. The D-flip-flop FF1 acts as a filter on the SYNC input that is driven by a SPAD and enables the non-PVT compensated ring oscillator. A 7-bit counter is used to reset the filter which in turn stops the clock generating oscillator.

The timing diagram of figure 5.5 illustrates the working operation. For glitchless clock generation, the following relation must be true:

$$\delta_{\text{fb}} < \tau/2$$

where δ_{fb} is the delay of the feedback loop from/to the output of the NAND-gate passing through the underflowing counter, and τ is the clock period. When the clock frequency reaches the gigaherzt range, $\tau/2$ approaches δ_{fb}. Fortunately, the feedback loop delay can be easily adjusted by selecting the proper feedback point in the delay line. Selecting odd or even taps, a delay values in the range of $[0, \tau/2]$ and $[\tau/2, \tau]$ respectively can be chosen. Note that in the preceeding discussion, the jitter of the SYNC signal and the generated clock was deliberately omitted for clarity. A safety margin is also required to cope with these signals uncertainties. The feedback loop delay mechanism is also suitable for this.

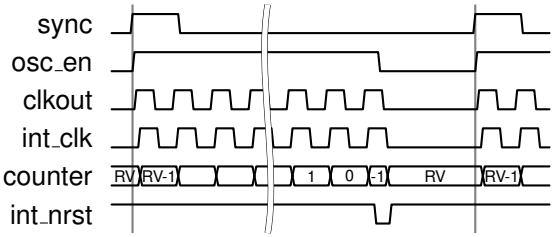


Figure 5.5: Timing diagram of the inner workings of Oscar. Note that the internal asynchronous reset, is active when the counter underflows.

The layout of the fixed frequency version of *Oscar* is shown in Figure 5.6. The three constituting elements are the SPAD, control logic, and fixed frequency ring oscillator. The SPAD has been described in section 2.1.1. Its active region is separated from the rest of the design by 10 µm in order to limit substrate noise injection. For the same reason, the ring oscillator has a triple guard ring

to capture substrate charges generated by the mass of switching inverter gates that form the oscillator. These measures may be relaxed in future implementations of *Oscar* in the interest of compactness. The controller, contains 46 digital cells and runs at 550 MHz. The total size of the cell is of 68 µm × 27 µm and could be compacted by at least 25 % with a smaller active area SPAD, shorter safeguard distances, and higher frequency oscillator.

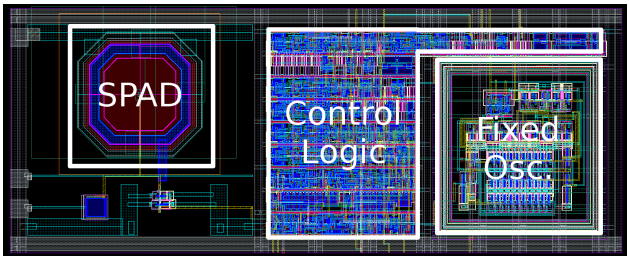


Figure 5.6: Layout of Oscar with fixed 550 MHz oscillator. Dimensions: 68 µm × 27 µm.

5.2.2 Variable Frequency *Oscar*

A detailed version of the schematic is shown in figure 5.7 in conjunction with figure 5.5, a timing diagram of the relevant signals in operation. Note here the SYNC signal selection in order to test with the accompanying SPAD as well as an electrical signal. The CLK-OUT output is the first tap of the delay chain in order to minimize the delay and jitter between the SYNC pulse and the first edge of the generated clock signal. Glitches at the output may arise due to the counter-reaction control loop delay. To avoid glitches, the 3-bit DELAY_SEL enables a fine selection of delays in the control loop by selecting several readily available shifted versions of the clock. This mechanism is especially necessary on the variable oscillator

version of *Oscar* because of the large difference between the oscillator's possible periods and the fixed delay of the counter-reaction loop from the counter to the oscillator's output.

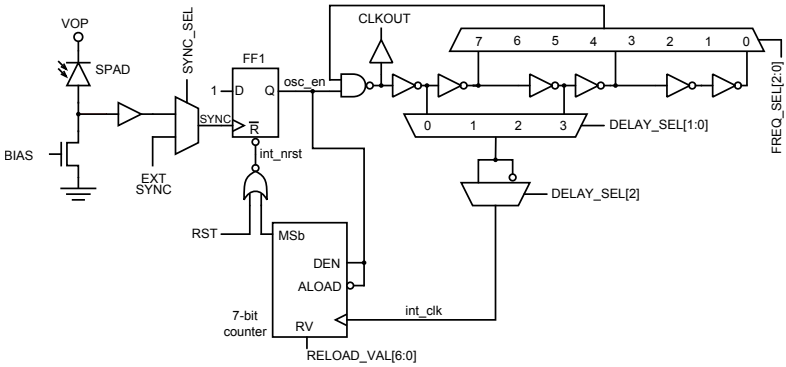


Figure 5.7: Detailed schematic of Oscar. The D-flip-flop acts as a filter on the sync input that is driven by a SPAD and enables the non-PVT compensated ring oscillator. A 7-bit counter is used to reset the filter which in turn stops the clock generating oscillator.

The layout of the variable frequency *Oscar* is shown in Figure 5.8. The three constituting elements are the SPAD, control logic, and variable ring oscillator. The SPAD has been described in section 2.1.1. A much smaller device could be beneficial in terms of area, noise, and afterpulsing, due to the reduced carriers involved in an avalanche. It's active region is separated from the rest of the design by 10 µm in order to limit substrate noise injection. For the same reason, the ring oscillator has a triple guard ring to capture substrate charges generated by the mass of switching inverter gates that form the oscillator. The controller occupying the space between the SPAD and the ring oscillator, contains 58 digital cells. The total size of the cell is of 53 µm × 66 µm.

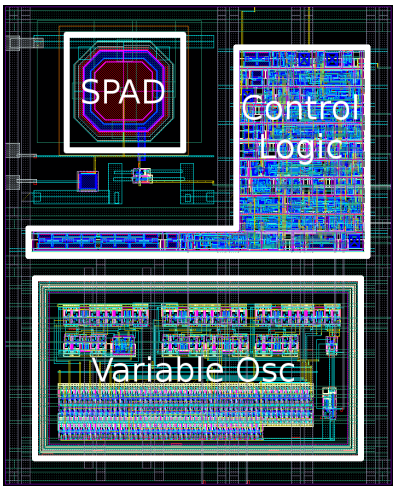


Figure 5.8: Layout of Oscar variable frequency oscillator. Dimensions: 53 µm × 66 µm

5.2.3 Metastability

Both fixed and variable oscillator implementations are subject to metastability issues on the filtering flip-flop FF1. In fact, if the recovery or removal times of this flip-flop are violated the system may become unstable or even oscillating. This is due to the fact that a metastable osc_en will propagate through the reset feedback loop to int_nrst. However, we force by design the SYNC signal to occur at a predefined interval δ_T. Therefore, for a given oscillator period δ_{osc} we choose a reload value RV such that

$$\delta_T - \delta_{recovery} - \delta_{prop} < RV \times \delta_{osc}$$

or

$$\delta_T + \delta_{removal} - \delta_{prop} > RV \times \delta_{osc}$$

where $\delta_{recovery}$ and $\delta_{removal}$ are the recovery and removal times of the reset signal with respect to the clock of the flip-flop. δ_{prop} is the propagation delay inherent to the feedback loop. Again, as discussed in section 5.2.1, the jitter of the SYNC signal and the generated clock impact metastability and an extra safety margin should be taken for this.

5.2.4 Skew and Jitter

A clock distribution network using *Oscar* must, like any other clock distribution network, control skew and jitter at all endpoints. As already mentioned, skew can be reduced to almost zero, thanks to the optical distribution approach. However, a mismatch due to technology variations might introduce a systematic offset between the leading edge of two *Oscar* generated clocks trees. Note that, in the scheme proposed. only the skew of the leading edge of the first clock cycle is important. The same is true for jitter. The jitter of the first clock edge here is dominated by the SPAD's jitter. In fact, the filter flipflop and NAND gate contributions are negligible. The sensor's jitter was not optimized in this design (400 ps for 90nm SPAD and 80 ps for an external 0.35 µm SPAD). For reference, commercial microprocessors have clock distribution jitter as low as a few picoseconds for multi-GHz clock frequencies at the cost of large silicon area.

5.3 Practical implementation of Oscar in VLSI

We present one possible application of the *Oscar* clocking scheme. In order to highlight the peculiarities of the proposed application, let us review those approaches that are relevant to it.

The distribution of clock signals, whether optical or electrical,

has a major role not only in synchronous systems [137] but also in systems with limited or localized synchronicity. An example of such systems are GALS systems [138], where important power savings are realized by creating localized clock domains and replacing a global clock with asynchronous data exchange protocols. In the GALS approach, a core optimization lies in the selection of a sweet spot frequency for the localized clock domains. Another issue is that of the selection of the proper data exchange protocol to minimize the area and power impact to the overall design. Reliable data transfer at high bandwidth between clock domains is addressed by several methods [138–140]. "Pausible Clocking" is one such mechanism of special interest to us. The idea is to "pause" the clock to allow safe latching of transmitted data between modules. The transmission can be done through FIFOs [141] or directly [142] but some synchronization is needed. These systems usually suffer from non-deterministic execution which complicates testing and validation [143].

The demand of widely specialized processors has lead to single-die, multi-die packaged, or multi-package heterogeneous systems. An example of the last category are coprocessor systems which were highly in vogue twenty-years ago [144, 145] though some more recent work revisits the paradigm [146–148]. ISEs can be seen as an evolution of coprocessors. While coprocessors expand processor functionality with datapath and control logic through a defined external interface, ISEs with CIUs, are only an addition to the processor's datapath. A careful choice of "accelerated" instructions is required and has been done manually until Clark *et al.* first demonstrated automatic selection [149]. Further research also confirmed the viability of automatic ISEs detection [150–155].

The application of *Oscar* proposed here builds on the experience of GALS, coprocessors, and ISEs while avoiding their limitations. The clocking circuit implementing these ideas was designed

into an integrated circuit and used to test a wide variety of trade-offs. The demonstrator chip was implemented in a standard 90 nm digital CMOS technology. Chips in this process can be thinned to several tens of micrometers, thus enabling the *Oscar* technology to be used in 3D stacks where the optical clock would be transmitted through chips.

The chip whose architecture is shown in figure 5.9, was fabricated in TSMC 90 nm CMOS technology. Two 32-bit OpenRISC processors were included along with a custom bus interface. The processors which will be described in more details in the following section, have 8 kB instruction cache and 8 kB data cache each.

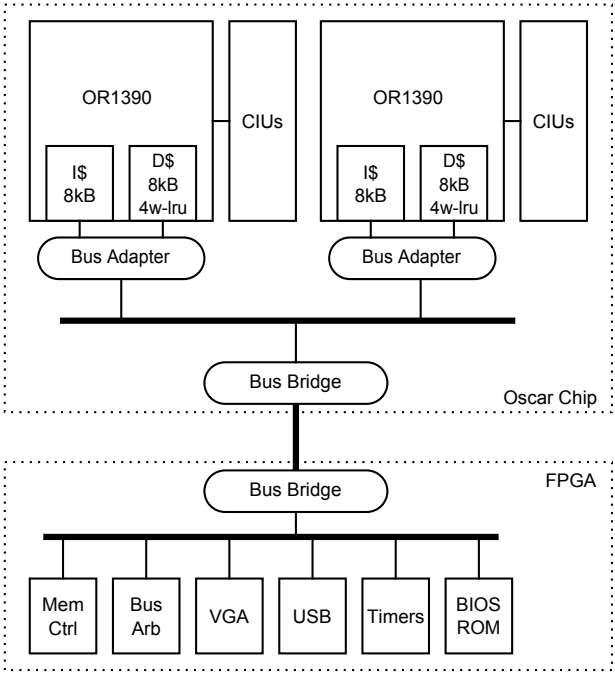


Figure 5.9: Architecture of the Oscar chip and system's peripherals.

Two sets of custom instructions were implemented. The first set was included in the left processor and was designed specifically to work with *Oscar*. The second set of custom instructions are designed to demonstrates Coherent DMA, Speculative DMA, Virtual Ways, and Way Stealing as presented by Kluter in [156–158].

The bus adapter allows the processor and bus to work at different clock frequencies. While we fixed the bus frequency at 50 MHz, the processors can run up to 260 MHz. In this pad-limited design, the bus bridge was designed to reduced the number of pads to interface with and maximize power supply pads. For power supply, 78 pads were used while 87 pads were used for data and control.

All the peripherals such as main memory, VGA and USB interfaces, and BIOS are implemented after the bridge in a FPGA. While this choice may not be optimal for performance, especially for the main memory, it is suitable for all the testing of *Oscar* where the programs and data used can fit the caches. The micrograph in figure 5.10 shows the pad-limited chip design. The die size is 3940 µm × 1875 µm for a total of 104 kGates.

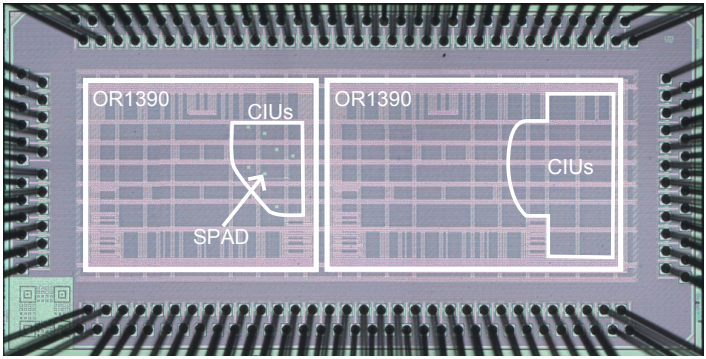


Figure 5.10: Micrograph of the Oscar chip fabricated in TSMC 90 nm CMOS technology. Die size: 3950 µm × 1875 µm. Gate Count: 104k.

5.3.1 Processor and Custom Instructions

The OpenRISC project [159] provides specifications of a free, open source 32/64-bit RISC/DSP architecture. It's conception is very well suited to embedded systems. The project provides the architecture, an open source implementation, and a complete software development kit. The OpenRISC 1000 instruction set-compliant processor used by Kluter in [156–158] was ported from FPGA fabric to ASIC. The ASIC derivation (OR1390) used in this work, was adapted for TSMC 90 nm CMOS semi-custom flow based on Low-V_{t} standard cells and memories. The processor has a 5-stage pipeline in-order architecture with 8 kilobytes 4-way set associative data cache and 8 kilobytes 2-way set associative instruction cache. Both caches use a LRU replacement policy. The custom instruction interface* allows multi-cycle custom instructions to be added to the processor. Figure 5.11 represents the usual timing diagram of a multi-cycle custom instruction call from the processor.

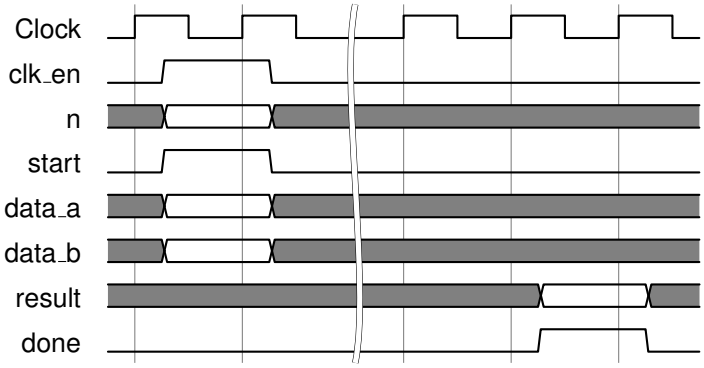


Figure 5.11: Timing diagram of a multi-cycle Altera NiosII compliant custom instruction.

*compliant with Altera Nios II [160]

5.3.2 Custom Instruction Units and Oscar

We manually implemented three CIUs in order to test the *Oscar* clocking scheme. Each of these three CIUs contains control logic in form of a Finite State Machine (FSM) to provide multi-cycle execution. The first CIU is a textbook implementation of a radix-1 non-performing restoring 16-bit integer divider. This radix-1 divider takes 17 cycles to complete. The second CIU is a classic multi-cycle 32-bit integer multiplier with 32-bit integer result. This CIU takes 36 cycles to complete. Finally we implemented a shifter that supports arithmetic and logic shifts as well as rotations. This shifter performs a single shift each cycle making its execution time in clock cycles dependent on the number of positions to shift.

The variable-cycle Custom Instructions (CIs) require that, for a fixed *Oscar* configuration, the done control signal be extended until the next synchronization timepoint. The added logic called done_wrapper is shown in figure 5.12. The start_wrapper was added in order to synchronize the start signal in the special case where the processor would also be clocked by *Oscar* and a CI call is not aligned on a synchronization boundary. Although not strictly necessary, these synchro wrappers make use of *Oscar* state information and are almost transparent in normal operation. Their asynchronous design introduces only combinatorial delay to the control signals path.

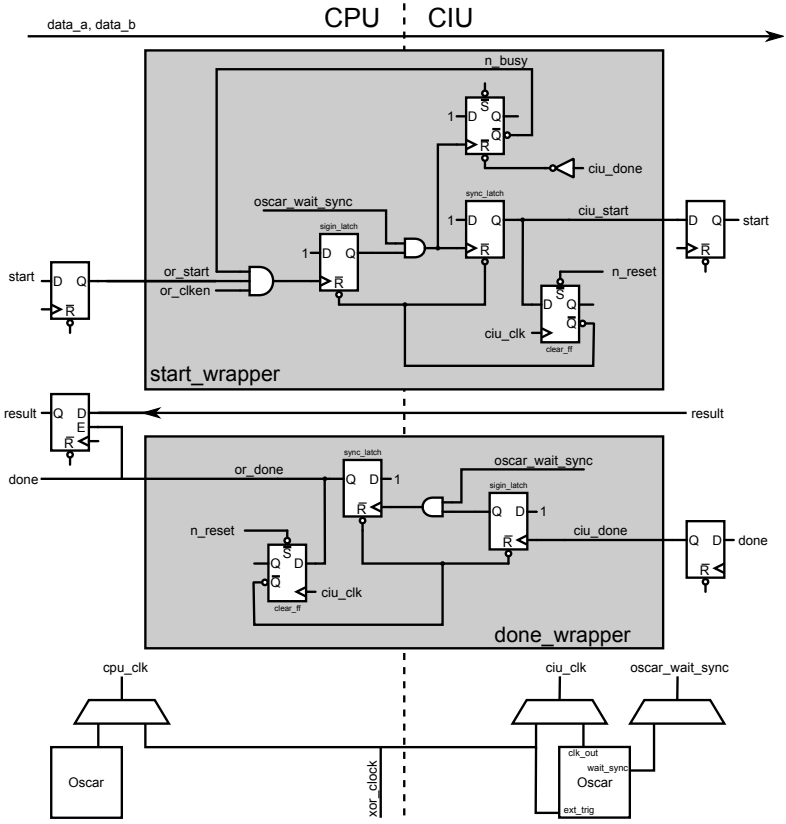


Figure 5.12: CI call synchronization logic is used to ensure the start and done control signals are extended to the synchronization timepoints.

5.4 Results

Validation of an ASIC design plays an important role in ensuring that design specifications be met before fabrication. Beside lengthy simulations at RTL or gate level, emulation is a key validation step. The process has been made popular by the wide adoption of FPGA platforms. Before presenting the results related to the use of *Oscar* clocking, we emphasize the validation of the system as a whole in the following section.

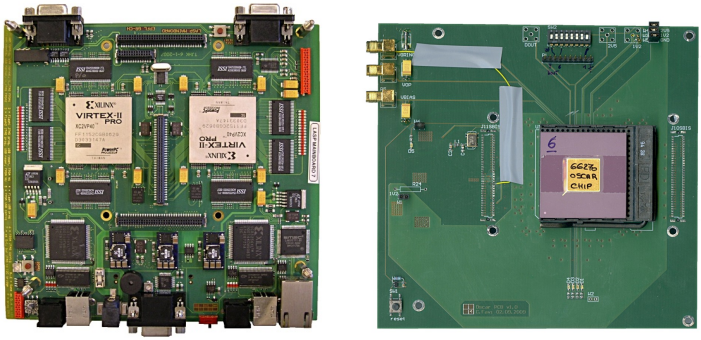
5.4.1 System Pre-validation

The system architecture described in figure 5.9 is based on a Chip–FPGA codesign. Validation of the whole system was performed on a dual FPGA board shown in figure 5.13(a). The FPGA on the left containing the memory controller, the bus arbiter, a VGA controller, the BIOS, a timers module, and the USB interface used for software transfer and configuration. The second FPGA held the same code used for the chip except for the technology specific parts (memories, flip-flops, and *Oscar*).

The bidirectional interface has been thoroughly tested between the two FPGAs. When moving to the *Oscar* chip daughterboard (fig. 5.13(b)), the differences in timing due to the high capacitative load of the connectors limit the maximum speed of the chip-to-mainboard communication to around 70 MHz.

5.4.2 Test setup and methodology

Functional tests were first performed with external clocking as opposed to *Oscar* clocking. Correct functionality of the CPUs and custom instructions were validated. Especially the CIs were thoroughly tested with all combinations of input values, when possible.



(a) The dual FPGA board used to validate the systems' architecture.

(b) Daughterboard holding the *Oscar* chip and interfacing to the FPGA board.

Figure 5.13: Motherboard and daughterboard used in the testing of the *Oscar* chip.

To measure the frequencies of the oscillators, we use the CPU frequency counter that basically counts the clock cycles in a millisecond. In order to measure the oscillator frequency f_{osc}, we set Oscar's sync at a frequency f_{sync}, successively increment the reload value RV, and record the reported frequency value. The maximum value approaches the real value and we see the following trend of reported frequencies:

$$f_{sync}, 2f_{sync}, \ldots, Nf_{sync}, \tfrac{N+1}{2}f_{sync}, \ldots, \tfrac{M}{2}f_{sync}, \tfrac{M+1}{3}f_{sync}, \ldots$$

This is easily explained by the fact that whenever

$$RV \times f_{sync} > k \times f_{osc},$$

we are crossing a synchronization time-point boundary and, therefore, the oscillator is stopped until the following sync arrives.

Power measurements were performed with a Tektronix TM502A current probe amplifier connected to a Picotech Picoscope 6403. A single measure is the average of 20 frames. A frame consists of 5 MSamples spanning 200 ms. The measurement precision, or reproducibility, was 1 % of the absolute value. We only sampled the core voltage (1.2 V) thus leaving out I/O power (2.5 V). Whenever we measured dynamic power, unused parts were deactivated through clock gating.

The optical setup consists of a 637 nm laser diode. The nominal frequency of the diode is 40 MHz. However, the laser diode controller was also clocked externally with a function generator at lower frequencies. The uncollimated laser beam was directly pointed to the surface of a 0.35 µm SPAD chip directly connect to the *Oscar* chip. The laser power was chosen to minimize pile-up effects. Figure 5.14 shows the optical setup used.

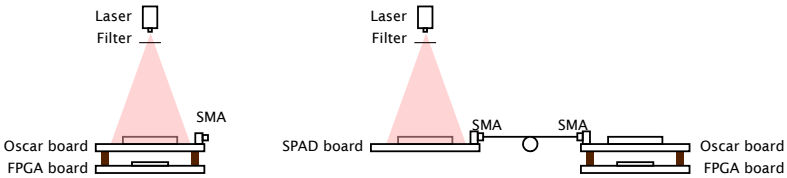


Figure 5.14: Left, the original setup to test the *Oscar* chip with the internal SPADs. Right, the setup used for the tests with electrical connection between a 0.35 µm SPAD chip and the *Oscar* chip.

All experiments were conducted at 20 °C and 1.2 V core voltage. In all the tests, the influence of the bus clock, fixed at 50 MHz, has been minimized. For example, the test operations are performed on cached data values or processor registers to prevent bus accesses besides the initial mandatory external memory fetches. This method maximizes power consumption and performance since a bus access would stall the processor for several cycles.

All chip control signals such as clocking muxes and *Oscar* parameters, were configured at run-time by the processor. The software was built with a customized toolchain based on GCC 3.4.4 in which custom instruction assembly opcodes were added. A JTAG-like interface is also available to set these parameters externally. Figure 5.15 show the clocking architecture in place to test oscar.

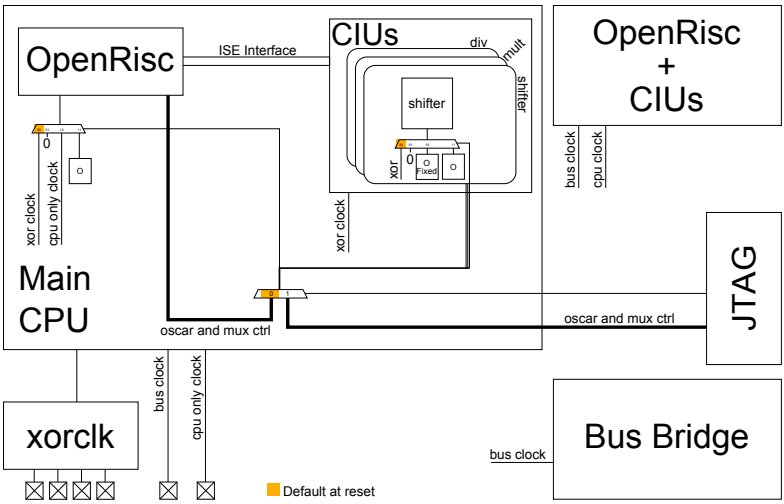


Figure 5.15: The clocking of the different parts of the chip can be selected either by the main CPU or externally by a JTAG-like interface.

5.4.3 Measurements

The clock frequencies of the variable oscillator range from 114 MHz to 534 MHz while the fixed frequency oscillator runs at 502 MHz. These measures vary within 5 % across different chips. The processor was validated to run up to 260 MHz.

Power measurements independent of *Oscar* clocking operation are reported in the following table:

Static Power	42	µW
Dynamic Power	0.24	mW/MHz

Static power includes the complete dual core system except IO power. Dynamic power for one CPU is measured with a tight loop of operations rearranging assembly code to prevent data dependency as much as possible. The maximum value is selected. Note that only the 1.2 V core power is reported in this measurement.

Three operations were tested: division, multiplication, and logic shift left. For a given operation, a loop of 10 000 iterations is run. The elapsed wall time for the loop execution is recorded in order to compute the Million of Operations Per Second (MOPS) figure. Depending on the operation, several ways of execution were tested. For the division, we used the software division available in the toolchain, the CIU clocked normally, and the CIU clocked with *Oscar*. For the multiplication, we ran the tests with the datapath's single-cycle multiplier, the CIU clocked normally, and the CIU clocked with *Oscar*. For the shift operation, we only compare the single-cycle datapath shifter with the CIU clocked with *Oscar*. In fact, for this operand-dependent variable-cycle instruction, the power is highly correlated with the operand value. Figures 5.16, 5.17 and 5.18 present power measurements versus million of operations per second (MOPS).

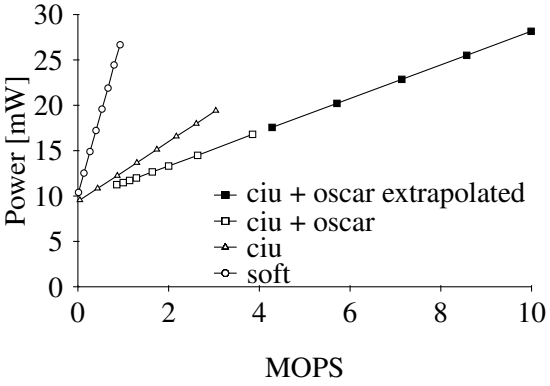


Figure 5.16: Power measurements of the division operation. Note the expected linear trend of each set of measurements. The software division is here compared to the CIU only, our system lacking dedicated hardware division.

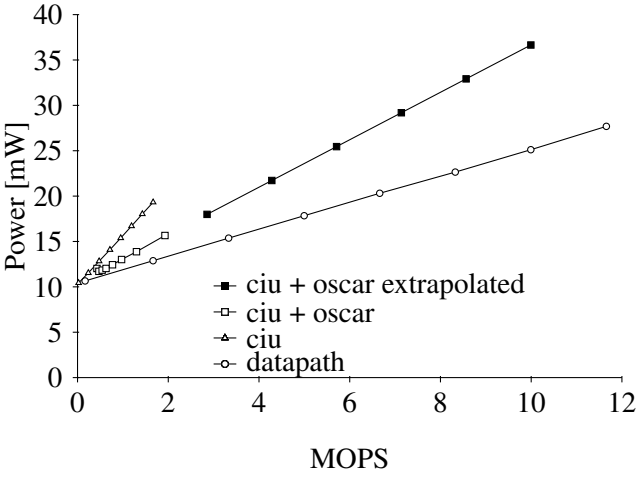


Figure 5.17: Power measurements of the multiplication operation. Note the expected linear trend of each set of measurements. The datapath multiplier is compared to the multi-cycle CIU.

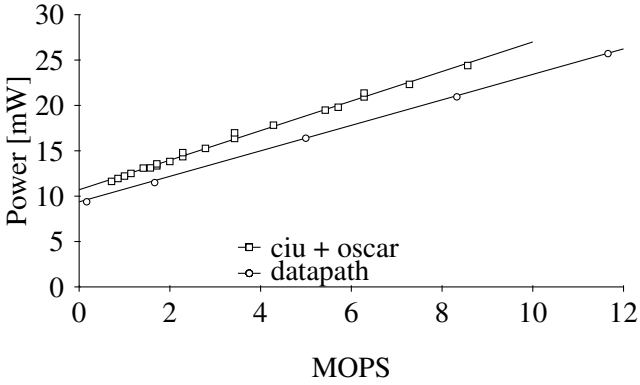


Figure 5.18: Power measurements of the logic shift left operation. The data dependent variable-cycle CIU clocked with *Oscar* is run with a variety of input values, however the trend of the curve is still linear. The internal datapath shifter is also shown as reference.

The offset of approximatively 10 mW is the sum of static power and dynamic power due to the bus circuitry running at 50 MHz.

To be fair when comparing the different implementations, we use the following well-known figure of merit:

$$\kappa = \text{performance}/(\text{power} \times \text{area})$$

Since our design is pad-limited, the area of the functional units, both CIUs and datapath, were recomputed separately. The constraints were 80 % cell area usage, 500 MHz clock for datapath unit, and 2 GHz clock for the CIUs.

The figure of merit for the multiplication and shift operations are reported in Table 5.1. Note that division is only represented with the CIU because of the lack of single-cycle datapath divider unit in our OpenRISC architecture.

Instruction	Perf/Power [MOPS/mW]	Area [10^{-3} mm^2]	Figure of Merit [MOPS/(mW·mm^2)]
Mult DP	0.677	14.558	46.543
Mult CIU	0.184	4.573	40.303
Mult CIU+Oscar	0.382	4.573	83.631
Shift DP	0.711	2.077	342.816
Shift CIU+Oscar	0.636	1.598	398.326
Div CIU	0.304	2.874	105.775
Div CIU+Oscar	0.539	2.874	187.543

Table 5.1: Figures of merit and area of CIUs and datapath units.

5.5 Discussion

The *Oscar* design shares some similarities with [161]. While applied to GALS designs, the digitally controlled clock multiplier of [161] also uses a gated ring-oscillator and counter. We differ from that design by being exclusively standard-cell based albeit being slightly less efficient in area, power and noise, because of the number of inverters and multiplexers used.

When compared to GALS designs, our clocking scheme is completely deterministic. In fact, our design is completely synchronous and all the known design verification rules still apply. For example, the synthesis constraints are set such that the CIU clocks frequencies are a multiple of the CPU clock. In this way, the synthesizer takes care of setup and hold time across time domains.

From the literature it is known that the use of photonics as means of clock distribution can only replace a small part of the total clock network power [119, 139]. The large capacitance to be switched is generally at the leaves of the clocking tree. The granularity of placing the optical-to-electrical converters is highly dependent of their area but also to the optical distribution means. For

example, if the sensors were extremely small, one could have optical clock latches. The optics required to efficiently distribute the light could be based on holography. However, we place ourselves in a larger-grain approach by only clocking a few *Oscar* units. Optical distribution could be done with fiber optics although for sake of simplicity we beamed a sufficiently powered laser over the entire surface of the chip.

From the result section, multi-cycle CIUs can have better power efficiency with *Oscar* clocking. The difference of power between operations with and without *Oscar* is due to the CPU frequency difference. In fact, for the same performance target (MOPS), the CPU frequency of the *Oscar* operated system is a fraction of that of a non-*Oscar* operated one. Furthermore some energy is wasted in the latter while the stalled-cpu waits for the CIU's result.

When comparing datapath single-cycle operations (multiplication and shift) with the *Oscar* enabled CIUs, we note that the former are more power efficient. However this comes at the cost of larger on die area. The introduction of the figure of merit tries arbitrarily to balance these trade-offs. The benefits of *Oscar* may not be completely exploited in the area unconstrained case however the simplicity of both the clocking circuitry and CIs allow fast design development. Finally note that these power results are independent of the use of an optical clock.

Although only the electric input of *Oscar* could be tested, the non-idealistic features of the SPADs are mitigated in this design in different ways. Reload time, after pulsing, and dark count are mitigated by the filtering inherent to the function of *Oscar*. The triggering window is purposely left small enough at the end of the clocking period so that spurious hits' impact is minimal. Any afterpulses are filtered by the SYNC flip-flop FF1 in figures 5.4 and 5.7. The jitter of the SPAD was not optimized, however it could be reduced easily by employing several sensors and OR-wiring their

outputs. The fixed frequency oscillator was meant to run at 1 GHz while variable oscillator would have selected frequencies between 200 MHz and 1.5 GHz. However a mistake in simulating the design yielded approximatevely half of these values. This should be fixed for a future design.

5.6 Conclusion

Single-Photon clocking as presented in this chapter shows some interesting applications. The simplicity of the design of *Oscar* itself and of the driven logic can be compelling in applications where power, performance, and area are serious constraints. Other applications could span from DSP to Network processors. In the security application context, *Oscar* could be used to generate a random-clock. In fact, when illuminated with non-coherent light, a SPAD produces uniformly spaced pulses (Poisson arrival times). This can be used to generate a truly random-clock with some simple filtering.

Also, the *Oscar* clocking scheme could be used in a 3D stack. Die thinning would be required for the optical clock to pass through to underlying dies. A discussion of die thinning can be found in [135]. Trade-offs between emitter power, die thinning and wavelength are required to be evaluated. *Oscar* clocking in this context could be achievable, however data-communication between dies should also be addressed.

We presented a clocking scheme based on a globally optically synchronized local oscillator named *Oscar*. It was applied to Instruction Set Extensions for an embedded processor using multi-cycle Custom Instruction Units. We presented an implementation of *Oscar* in 90 nm CMOS with the infrastructure around it. We discussed the power measurements results as a trade-off between performance, power and area. We briefly commented on similarities

and differences compared to Globally Asynchronous Locally Synchronous systems. Finally, we presented some applications of this clocking scheme whether for security as truly random-clock generation or for 3D integration.

6 OUTLOOK

SEVERAL other applications that can take advantage of highly integrated standard CMOS SPADs are of great interest. A few of them are mentionned in the following paragraphs.

Fluorescence is the result of spontaneous emission in fluorophores (fluorescent molecules) after absorption of light or other electromagnetic radiations. The lifetime of the fluorescence is a measure linked to the exponential decay rate of this process. The decay rate of fluorescence is highly dependent on environmental parameters. For example, in biological applications, these parameters would be concentration of ions (Na^+, Mg^{++}, Ca^{++}), oxygen concentration, or pH value [105]. Specifically designed fluorescent dyes have been developed where the decay rate changes as a function of the environmental parameter chosen. Hence by measuring the emission decay on a per pixel basis, one can create an image of these parameters. This technique is known as FLIM.

TCSPC can be used to precisely measure such decay; SPAD-array based prototypes used for FLIM applications have been shown by Gersbach *et al.* [162–165]. The particularity of the sensor chip described in [164] is the integration of a per pixel time-to-digital converter that can be used for time-correlated or time-uncorrelated imaging. Following the same idea, [112] presents, to date, the largest array of SPADs with on-pixel TDCs. The 160×128 array chip is shown in figure 6.1. The chip has been bonded directly on the board to limit inductive and capacitive effects of packaging on the 160 Mbps readout lanes. Impedance controlled and matched length lines were used to connect the 256 data lines* to two Xilinx

*two times 128 lines in fact: the chip's readout is split vertically in two parts.

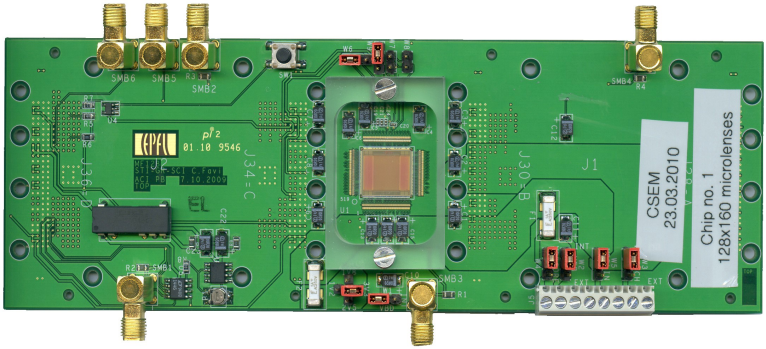


Figure 6.1: Prototype of 128×160 130 nm CMOS SPAD array used for instance in FLIM applications [112] designed by us. This daughter-board connects the sensor chip to two Virtex-4 devices on the motherboard with 320 length-matched impedance-controlled data lines for high-speed operation.

Virtex 4 FPGAs on a dedicated motherboard were all signal processing is performed before data transfer to a computer.

Several other applications of SPAD-based imaging techniques exist. In the bioimaging area, Förster Resonance Energy Transfer (FRET) is described in [105] while Fluorescence Correlation Spectroscopy (FCS) and FLIM are described in [166]. Positron emission tomography (PET) could take advantage of the high integration attainable with SPADs if faster scintillators are developed in the future [167].

High-energy physics has been the driver for fast, low jitter, high sensitivity detectors. In fact most of the research on photodiodes emerged from the need of detecting particles' induced photons arrival time with very precise timing. As direct detection of the particles is not always achievable, practicality imposed the use of scintillator-optodetector pairs. While PMTs, MCP, and APDs in

sub-Geiger mode are still used in this field [168, 169], Geiger mode photodiodes have been used as well [170].

In [103], a gamma, X-ray, and high energy proton radiation-tolerant SPAD has been developed for space applications. The 32×32 imager was designed to detect the atmospheric oxygen emission due to oxygen recombination, also known as Earth's airglow. Satellites use this phenomenon to infer their position with respect to earth in order to maintain geostationary orbit. The choice of a standard CMOS technology was driven by size and weight for inclusion in a low-cost micro-satellite. This solution could be further miniaturized with key benefits being lower DCR, lower dead time, lower afterpulsing, and lower optical cross-talk [171].

A different kind of application is that of quantum cryptography. A good overview of the field is given by Gisin *et al.* in [172]. Random-number generation (RNG) is of particular interest for cryptography and secure communication. The main difficulty in RNGs is to generate truly random numbers and doing so at high frequency rates. Beamsplitting random number generators were shown in [173, 174] for example. These systems take advantage of the corpuscular nature of photons by forcing the photons to choose a path at a Y intersection (beamsplitter). The length of the two paths is different and a single-photon detector at each end reveals which path was taken. Commercial products such as ID Quantique's use semi-transparent mirrors to achieve the same goals [175]. These methods generally reach a generation rate of circa 4 Mbit/s. In [176], the time between the arrival of two photons is measured with a restartable clock. The random bits are generated by comparing two consecutive measurements, t_1 and t_2, $t_1 < t_2$ yielding a 0, $t_1 > t_2$ a 1, and no bits are generated when $t_1 = t_2$. The rate of random-number generation achieved in that work is 1 Mbit/s. In [177] the output of a InGaAs APD illuminated with a attenuated distributed feedback continuous wave laser is combined with a local 1.03 GHz

gating clock. After filtering, the position of the detected photons is tagged to produce the random bitstream according to whether the photon arrived in an even ('1') or an odd ('0') clock cycle. This method leads to a random-number generation rate of 4.01 Mbit/s.

This work has no pretension to being exhaustive in listing all the possible applications of SPADs in digital CMOS technology. The practicality of few proposed applications is within the limits of today's technology. The concepts illustrated by these applications may be amenable to be employed in other contexts. With advances in integration of SPADs with lower noise, lower jitter, and lower dead time, the trade-off taken today will need to be reassessed. The basic assumption of this work is that high integration of SPAD technology is available. To conclude this dissertation, we will present future work related to our contributions in chapters 2–5.

Single-photon detection still needs to be perfected in deep submicron technology. While SPADs in 90 nm and 130 nm have been demonstrated, much effort still needs to be profused in order to fine tune performance. The main challenges in deep sub-micron technology are premature edge breakdown and tunneling effects, both of which affect noise. High doping concentrations and difficulties in controlling the fabrication process contribute to these effects. While devices built in dedicated processes will arguably have better performance than the ones built in standard processes, the intimate knowledge of the process in the second case has enabled us to achieve excellent results. The use of shallow trench isolation (STI), combined with controlled doping gradients at its surface shown in [165], or the use of an optional deep N implant to form a triple-well twin-tub structure shown in [178], are two good examples of DCR reduction.

In the field of short range optical communications, as described in chapter 3, we demonstrated theoretical channel capacities in excess of 100 Gbps. However for a practical implementation several issues are still to be solved. First, source coding is required for reliable communication over a noisy channel. On both the transmitter and the receiver sides, an implementation of the required decoding logic is needed in order to assess area and power utilization. Second, the proposed modulation scheme PPM, requires precise clock synchronization. The effects of jitter and clock mismatch on noise need to be bounded and controlled. Finally, a practical implementation of a packaged-aware optical channel still needs to be shown. We believe that the *all digital* nature of the proposed system has interesting advantages over traditional methods. However, at the time of this writing, we cannot yet make a strong argument in favor or against of the proposed approach. The time-to-digital converter in section 3.3 was designed in FPGA as a proof-of-concept for a demodulator. Although its single-shot performance did not reach the desired figures for PPM demodulation, it can still reach excellent resolutions in situations where averaging is possible. FPGA based TDCs are increasingly important in initiatives such as OpenPET. We believe that further development of this technology is essential.

Future work on SPAD-based imagers may go toward integration of larger array of pixels, increased timing accuracy, faster frame readout, in-pixel or on-chip processing, or any combinations thereof [179]. Inexhaustive as this list might be, we believe that standard CMOS single-photon imaging is becoming a viable solution for applications such as time-of-flight cameras and FLIM/FRET. As larger arrays are constructed and higher frame rates targeted, managing the large amount of data generated will become critical. Whether processed on-chip or off-chip, high bandwidth transfers will be required and new readout strategies enacted. Also light concentration techniques such as microlenses need to be perfected

to reclaim inactive areas due to SPAD guard ring and in-pixel circuitry.

Optical clock distribution has the potential of providing effortless clock distribution over a large area. The granularity at which optical-to-electrical conversion should be applied has to be carefully chosen. In fact, the proposed system can only ensure clock synchronization between domains when the optical signal triggers the oscillators. Control of the jitter of the generated clocks still needs much attention. The implementation shown in chapter 5 uses a simple inverter-delay ring-oscillator. The use of differential signaling and capacitance-controlled delays would lead to a denser and more power efficient oscillator. Working in photon starved regime is required to prevent pile-up effects. If the system becomes highly integrated, this problem also needs to be addressed.

To summarize, the main contributions of this thesis are threefold: single-photon CMOS communication paradigms, single-photon processing and readout techniques, single-photon clocking and synchronization methods. Single-photon communication was achieved using a combination of SPADs and ultra-fast TDCs in a pulse position modulation scheme. In this context, theoretical channel capacity limits in the presence of noise and other non-idealities typical of SPADs were derived; a TDC with 17ps resolution was demonstrated in a standard FPGA fabric. To the best of our knowledge, at the time of its writing, this is the highest reported resolution for a TDC of this kind. Single-photon processing and readout was achieved in several technologies, focusing on image sensor design, whereby massive parallel architectures were studied and implemented in CMOS. Single-photon clocking and synchronization was demonstrated allowing potential zero skew systems irrespective of the chip area. The power benefits of this approach in embedded systems with instruction set extensions are particularly interesting. These contributions make this thesis worthwhile for further

studies in the field of embedded design and optimization as well as in the fields of supercomputing and multi-core VLSI design.

LIST OF ACRONYMS

TCSPC Time-Correlated Single-Photon Counting 69

TDC Time-to-Digital Converter 26

TOF Time-of-Flight 70

TUPC Time-Uncorrelated Photon Counting 67

VLSI Very Large Scale Integration 1

BIBLIOGRAPHY

[1] A. Rochas, M. Gani, B. Furrer, P. A. Besse, R. S. Popovic, G. Ribordy, and N. Gisin, "Single photon detector fabricated in a complementary metal–oxide–semiconductor high-voltage technology," *Review of Scientific Instruments*, vol. 74, no. 7, pp. 3263–3270, 2003. [Online]. Available: http://link.aip.org/link/?RSI/74/3263/1

[2] Intel Corp., "Tera-scale silicon photonics research," Website, 2010, http://techresearch.intel.com/articles/Tera-Scale/1419.htm.

[3] IBM Corp., "Photonics & optoelectronics research," Website, 2010, http://www.zurich.ibm.com/st/photonics/interconnects.html.

[4] F. Doany, C. Schow, C. Baks, D. Kuchta, P. Pepeljugoski, L. Schares, R. Budd, F. Libsch, R. Dangel, F. Horst, B. Offrein, and J. Kash, "160Gb/s bidirectional polymer-waveguide board-level optical interconnects using CMOS-based transceivers," *IEEE Transactions on Advanced Packaging*, vol. 32, no. 2, pp. 345–359, May 2009.

[5] A. Rochas, "Single photon avalanche diodes in CMOS technology," Ph.D. dissertation, EPF-Lausanne, Switzerland, 2003.

[6] C. Niclass, A. Rochas, P.-A. Besse, R. S. Popovic, and E. Charbon, "CMOS imager based on single photon avalanche diodes," in *The 13th International Conference on Solid-State Sensors, Actuators and Microsystems. Digest of Technical Papers. TRANSDUCERS '05.*, vol. 1, Jun. 2005, pp. 1030–1034.

[7] C. Niclass, M. Sergio, and E. Charbon, "A single photon avalanche diode array fabricated in 0.35-um CMOS and based on an event-driven readout for TCSPC experiments," *Advanced Photon Counting Techniques*, vol. 6372, no. 1, 2006. [Online]. Available: http://link.aip.org/link/?PSI/6372/63720S/1

[8] M. A. Marwick and A. G. Andreou, "Single photon avalanche photodetector with integrated quenching fabricated in TSMC 0.18 um 1.8 V CMOS process," *Electronics Letters*, vol. 44, no. 10, pp. 643–644, May 2008.

[9] N. Faramarzpour, M. J. Deen, S. Shirani, and Q. Fang, "Fully integrated single photon avalanche diode detector in standard CMOS 0.18-um technology," *IEEE Transactions on Electron Devices*, vol. 55, no. 3, pp. 760–767, Mar. 2008.

[10] H. Finkelstein, M. J. Hsu, and S. C. Esener, "STI-bounded single-photon avalanche diode in a deep-submicrometer CMOS technology," *IEEE Electron Device Letters*, vol. 27, no. 11, pp. 887–889, Nov. 2006.

[11] C. Niclass, M. Gersbach, R. Henderson, L. Grant, and E. Charbon, "A single photon avalanche diode implemented in 130-nm CMOS technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 13, no. 4, pp. 863–869, Jul. 2007.

[12] M. Gersbach, C. Niclass, E. Charbon, J. Richardson, R. Henderson, and L. Grant, "A single photon detector implemented in a 130nm CMOS imaging process," in *38th European Solid-State Device Research Conference*, Sep. 2008, pp. 270–273.

[13] M. Gersbach, J. Richardson, E. Mazaleyrat, S. Hardillier, C. Niclass, R. Henderson, L. Grant, and E. Charbon, "A low-noise single-photon detector implemented in a 130 nm CMOS imaging process," *Solid-State Electronics*, vol. 53, no. 7, pp. 803–808, 2009, papers Selected from the 38th European Solid-State Device Research Conference - ESSDERC'08. [Online]. Available: http://www.sciencedirect.com/science/article/B6TY5-4W7B54W-1/2/e22e411a0ed8a5a32fd2791600c9d638

[14] M. A. Karami, M. Gersbach, and E. Charbon, "A new single-photon avalanche diode fabricated in 90nm standard CMOS technology," in *SPIE Optics and Photonics*, 2010.

[15] M. A. Karami, M. Gersbach, H.-J. Yoon, and E. Charbon, "A new single-photon avalanche diode in 90nm standard CMOS technology," *Opt. Express*, vol. 18, no. 21, pp. 22 158–22 166, 2010. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-18-21-22158

[16] C.-T. Sah, *Fundamentals of solid-state electronics*. World Scientific, 1991, 422–423.

[17] S. X. Jin, J. Shakya, J. Y. Lin, and H. X. Jiang, "Size dependence of III-nitride microdisk light-emitting diode characteristics," in *Applied Physics Letters*, 2001, vol. 78, pp. 3532–3534.

[18] H. X. Zhang, E. Gu, C. W. Jeon, M. D. Gong, Z.and Dawson, M. A. Neil, and P. M. French, "Microstripe-array InGaN light-emitting diodes with individually addressable elements," in *Photonics Technology Letters*. IEEE, Aug. 2006, vol. 18, pp. 1681–1683.

[19] C.-W. Jeon, H. W. Choi, E. Gu, and M. D. Dawson, "High-density matrix-addressable AlInGaN-based 368-nm microarray light-emitting diodes," *IEEE Photonics Technology Letters*, vol. 16, no. 11, pp. 2421–2423, Nov. 2004.

[20] H. W. Choi, C. Liu, E. Gu, G. McConnell, J. M. Girkin, I. M. Watson, and M. D. Dawson, "GaN micro-light-emitting diode arrays with monolithically integrated sapphire microlenses," *Applied Physics Letters*, vol. 84, no. 13, pp. 2253–2255, 2004. [Online]. Available: http://link.aip.org/link/?APL/84/2253/1

[21] Z. Gong, H. X. Zhang, E. Gu, C. Griffin, M. D. Dawson, V. Poher, G. Kennedy, P. M. W. French, and M. A. A. Neil, "Matrix-addressable micropixellated InGaN light-emitting diodes with uniform emission and increased light output," *IEEE Transactions on Electron Devices*, vol. 54, no. 10, pp. 2650–2658, Oct. 2007.

[22] B. R. Rae, C. Griffin, K. R. Muir, J. M. Girkin, E. Gu, D. R. Renshaw, E. Charbon, M. D. Dawson, and R. K. Henderson, "A microsystem for time-resolved fluorescence analysis using CMOS single-photon avalanche diodes and micro-LEDs," in *IEEE International Solid-State Circuits Conference. Digest of Technical Papers.*, Feb. 2008, pp. 166–603.

[23] W. L. Ng, M. A. Lourenco, R. M. Gwilliam, S. Ledain, G. Shao, and K. P. Homewood, "An efficient room-temperature silicon-based light-emitting diode," *Nature*, vol. 410, pp. 192–194, Mar. 2001.

[24] M. A. Green, J. Zhao, A. Wang, P. J. Reece, and M. Gal, "Efficient silicon light-emitting diodes," *Nature*, vol. 412, pp. 805–808, Aug. 2001.

[25] O. Boyraz and B. Jalali, "Demonstration of a silicon Raman laser," *Opt. Express*, vol. 12, no. 21, pp. 5269–5273, 2004. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-12-21-5269

[26] H. Rong, A. Liu, R. Jones, O. Cohen, D. Hak, A. Fang, and M. Paniccia, "An all-silicon Raman laser," *Nature*, vol. 433, pp. 292–294, Jan. 2005.

[27] H. Rong, R. Jones, A. Liu, O. Cohen, D. Hak, A. Fang, and M. Paniccia, "A continuous-wave Raman silicon laser," *Nature*, vol. 433, pp. 725–728, Feb. 2005.

[28] A. W. Fang, H. Park, O. Cohen, R. Jones, M. J. Paniccia, and J. E. Bowers, "Electrically pumped hybrid AlGaInAs-silicon evanescent laser," *Opt. Express*, vol. 14, no. 20, pp. 9203–9210, 2006. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-14-20-9203

[29] M. A. Green and M. J. Keevers, "Optical properties of intrinsic silicon at 300 K," *Prog. Photovolt: Res. Appl.*, vol. 3, no. 3, pp. 189–192, 1995. [Online]. Available: http://dx.doi.org/10.1002/pip.4670030303

[30] S.-H. Hsu, C. Panja, M. Lee, P.-N. Dong, and J. Chan, "Photonic bus with multi-channel selection on a SOI chip," in *Optical Fiber Communication Conference Technical Digest. OFC/NFOEC*, vol. 1, Mar. 2005.

[31] T. Baehr-Jones, M. Hochberg, and A. Scherer, "All-optical modulation in a silicon waveguide based on a single-photon process," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 14, no. 5, pp. 1335–1342, Sep. 2008.

[32] S.-H. Hsu, "A 5-um-thick SOI waveguide with low birefringence and low roughness and optical interconnection using high numerical aperture fiber," *IEEE Photonics Technology Letters*, vol. 20, no. 12, pp. 1003–1005, Jun. 2008.

[33] Soitec, "SOI products," Website, 2010, http://www.soitec.com/.

[34] G. V. Treyz, P. G. May, and J.-M. Halbout, "Silicon optical modulators at 1.3 μm based on free-carrier absorption," *IEEE Electron Device Letters*, vol. 12, no. 6, pp. 276–278, Jun. 1991.

[35] U. Fischer, T. Zinke, B. Schuppert, and K. Petermann, "Single-mode optical switches based on SOI waveguides with large cross-section," *Electronics Letters*, vol. 30, no. 5, pp. 406–408, Mar. 1994.

[36] C. K. Tang and G. T. Reed, "Highly efficient optical phase modulator in SOI waveguides," *Electronics Letters*, vol. 31, no. 6, pp. 451–452, Mar. 1995.

[37] A. Huang, C. Gunn, G.-L. Li, Y. Liang, S. Mirsaidi, A. Narashimha, and T. Pinguet, "A 10Gb/s photonic modulator and WDM MUX/DEMUX integrated with electronics in 0.13 μm SOI CMOS," *IEEE International Solid-State Circuits. Conference Digest of Technical Papers*, pp. 922–929, Feb. 2006.

[38] A. Liu, L. Liao, D. Rubin, H. Nguyen, B. Ciftcioglu, Y. Chetrit, N. Izhaky, and M. Paniccia, "High-speed optical modulation based on carrier depletion in a silicon waveguide," *Opt. Express*, vol. 15, no. 2, pp. 660–668, 2007. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-15-2-660

[39] L. Liao, A. Liu, J. Basak, H. Nguyen, M. Paniccia, D. Rubin, Y. Chetrit, R. Cohen, and N. Izhaky, "40 Gbit/s silicon optical modulator for highspeed applications," *Electronics Letters*, vol. 43, no. 22, Oct. 2007.

[40] W. M. Green, M. J. Rooks, L. Sekaric, and Y. A. Vlasov, "Ultra-compact, low RF power, 10 Gb/s silicon Mach-Zehnder modulator," *Opt. Express*, vol. 15, no. 25, pp. 17 106–17 113,

2007. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-15-25-17106

[41] D. Marris-Morini, L. Vivien, F. J. M., E. Cassan, P. Lyan, and S. Laval, "Low loss and high speed silicon optical modulator based on a lateral carrier depletion structure," *Opt. Express*, vol. 16, no. 1, pp. 334–339, 2008. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-16-1-334

[42] K. M. Brown, "System in package 'the rebirth of SIP'," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, Oct. 2004, pp. 681–686.

[43] N. Miura, D. Mizoguchi, T. Sakurai, and T. Kuroda, "Analysis and design of inductive coupling and transceiver circuit for inductive inter-chip wireless superconnect," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 4, pp. 829–837, Apr. 2005.

[44] R. J. Drost, R. D. Hopkins, R. Ho, and I. E. Sutherland, "Proximity communication," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1529–1535, Sep. 2004.

[45] A. Narasimha, B. Analui, E. Balmater, A. Clark, T. Gal, D. Guckenberger, S. Gutierrez, M. Harrison, R. Ingram, R. Koumans, D. Kucharski, K. Leap, Y. Liang, A. Mekis, S. Mirsaidi, M. Peterson, T. Pham, T. Pinguet, D. Rines, V. Sadagopan, T. J. Sleboda, D. Song, Y. Wang, B. Welch, J. Witzens, S. Abdalla, S. Gloeckner, and P. De Dobbelaere, "A 40 Gb/s QSFP optoelectronic transceiver in a 0.13 µm CMOS silicon-on-insulator technology," in *Conference on Optical Fiber communication/National Fiber Optic Engineers Conference. OFC/NFOEC*, Feb. 2008, pp. 1–3.

[46] S. Assefa, F. Xia, and Y. A. Vlasov, "Reinventing germanium avalanche photodetector for nanophotonic on-chip optical interconnects," *Nature*, vol. 464, pp. 80–84, Mar. 2010.

[47] J. Pierce, "Optical channels: Practical limits with photon counting," *IEEE Transactions on Communications*, vol. 26, no. 12, pp. 1819–1821, Dec. 1978.

[48] H. M. H. Shalaby, "Performance analysis of optical synchronous CDMA communication systems with PPM signaling," *IEEE Transactions on Communications*, vol. 43, no. 234, pp. 624–634, Feb. 1995.

[49] K. Kiasaleh, "Turbo-coded optical PPM communication systems," *Journal of Lightwave Technology*, vol. 16, no. 1, pp. 18–26, Jan. 1998.

[50] T. Ohtsuki and J. M. Kahn, "BER performance of turbo-coded PPM CDMA systems on optical fiber," *Journal of Lightwave Technology*, vol. 18, no. 12, pp. 1776–1784, Dec. 2000.

[51] B. Rae, J. McKendry, Z. Gong, E. Gu, D. Renshaw, M. Dawson, and R. Henderson, "A 200 MHz 300 ps 0.5 pJ/ns optical pulse generator array in 0.35 um CMOS," in *IEEE International Solid-State Circuits Conference. Digest of Technical Papers.*, Feb. 2010, pp. 322–323.

[52] J. McKendry, B. Rae, Z. Gong, K. Muir, B. Guilhabert, D. Massoubre, E. Gu, D. Renshaw, M. Dawson, and R. Henderson, "Individually addressable AlInGaN micro-LED arrays with CMOS control and subnanosecond output pulses," *IEEE Photonics Technology Letters*, vol. 21, no. 12, pp. 811–813, Jun. 2009.

[53] J. Abshire, "Performance of OOK and low-order PPM modulations in optical communications when using APD-based receivers," *IEEE Transactions on Communications*, vol. 32, no. 10, pp. 1140–1143, Oct. 1984.

[54] S. Kojima, "Demodulation method and demodulator of pulse-edge shifted pulse," *US Patent*, no. US 2010/0054370 A1, 2008.

[55] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[56] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, 1960. [Online]. Available: http://www.jstor.org/stable/2098968

[57] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: turbo-codes," *IEEE Transactions on Communications*, vol. 44, no. 10, pp. 1261–1271, Oct. 1996.

[58] R. McEliece, "Practical codes for photon communication," *IEEE Transactions on Information Theory*, vol. 27, no. 4, pp. 393–398, Jul. 1981.

[59] J. Massey, "Capacity, cutoff rate, and coding for a direct-detection optical channel," *IEEE Transactions on Communications*, vol. 29, no. 11, pp. 1615–1621, Nov. 1981.

[60] J. Liu, "Reliability of quantum-mechanical communication systems," *IEEE Transactions on Information Theory*, vol. 16, no. 3, pp. 319–329, May 1970.

[61] J. Pierce, E. Posner, and E. Rodemich, "The capacity of the photon counting channel," *IEEE Transactions on Information Theory*, vol. 27, no. 1, pp. 61–77, Jan. 1981.

[62] S. Butman, J. Katz, and J. Lesh, "Bandwidth limitations on noiseless optical channel capacity," *IEEE Transactions on Communications*, vol. 30, no. 5, pp. 1262–1264, May 1982.

[63] J. Lesh, "Capacity limit of the noiseless, energy-efficient optical PPM channel," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 546–548, Apr. 1983.

[64] F. M. Davidson and X. Sun, "Gaussian approximation versus nearly exact performance analysis of optical communication systems with PPM signaling and APD receivers," *IEEE Transactions on Communications*, vol. 36, no. 11, pp. 1185–1192, Nov. 1988.

[65] F. M. Davidson and X. Sun, "Slot clock recovery in optical PPM communication systems with avalanche photodiode photodetectors," *IEEE Transactions on Communications*, vol. 37, no. 11, pp. 1164–1172, Nov. 1989.

[66] H. Sugiyama and K. Nosu, "MPPM: a method for improving the band-utilization efficiency in optical PPM," *Journal of Lightwave Technology*, vol. 7, no. 3, pp. 465–472, Mar. 1989.

[67] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, Jan. 1972.

[68] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, Jul. 1972.

[69] B. K. Swann, B. J. Blalock, L. G. Clonts, D. M. Binkley, J. M. Rochelle, E. Breeding, and K. M. Baldwin, "A 100-ps time-resolution CMOS time-to-digital converter for positron emission tomography imaging applications," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 11, pp. 1839–1852, Nov. 2004.

[70] K. Karadamoglou, N. P. Paschalidis, E. Sarris, N. Stamatopoulos, G. Kottaras, and V. Paschalidis, "An 11-bit high-resolution and adjustable-range CMOS time-to-digital converter for space science instruments," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 1, pp. 214–222, Jan. 2004.

[71] H. Matsumoto, O. Sasaki, K. Anraku, and M. Nozaki, "Low power high resolution TDC with fast data conversion for balloon-borne experiments," *IEEE Transactions on Nuclear Science*, vol. 43, no. 4, pp. 2195–2198, Aug. 1996.

[72] K. Maatta and J. Kostamovaara, "A high-precision time-to-digital converter for pulsed time-of-flight laser radar applications," *IEEE Transactions on Instrumentation and Measurement*, vol. 47, no. 2, pp. 521–536, Apr. 1998.

[73] F. Bigongiari, R. Roncella, R. Saletti, and P. Terreni, "A 250-ps time-resolution CMOS multihit time-to-digital converter for nuclear physics experiments," *IEEE Transactions on Nuclear Science*, vol. 46, no. 2, pp. 73–77, Apr. 1999.

[74] J. Kalisz, R. Szplet, J. Pasierbinski, and A. Poniecki, "Field-programmable-gate-array-based time-to-digital converter with 200-ps resolution," *IEEE Transactions on Instrumentation and Measurement*, vol. 46, no. 1, pp. 51–55, Feb. 1997.

[75] R. Szplet, J. Kalisz, and R. Szymanowski, "Interpolating time counter with 100 ps resolution on a single FPGA device," *IEEE Transactions on Instrumentation and Measurement*, vol. 49, no. 4, pp. 879–883, Aug. 2000.

[76] M. S. Andaloussi, M. Boukadoum, and E. M. Aboulhamid, "A novel time-to-digital converter with 150 ps time resolution and 2.5 ns pulse-pair resolution," *The 14th International Conference on Microelectronics*, pp. 123–126, Dec. 2002.

[77] J. Wu, Z. Shi, and I. Y. Wang, "Firmware-only implementation of time-to-digital converter (TDC) in field-programmable gate array (FPGA)," *IEEE Nuclear Science Symposium Conference Record*, vol. 1, pp. 177–181, Oct. 2003.

[78] J. Song, Q. An, and S. Liu, "A high-resolution time-to-digital converter implemented in field-programmable-gate-arrays," *IEEE Transactions on Nuclear Science*, vol. 53, no. 1, pp. 236–241, Feb. 2006.

[79] J. Wu and Z. Shi, "The 10-ps wave union TDC: Improving FPGA TDC resolution beyond its cell delay," in *IEEE Nuclear Science Symposium Conference Record*, 2008, pp. 3440–3446.

[80] S. Junnarkar, P. O'Connor, and R. Fontaine, "FPGA based self calibrating 40 picosecond resolution, wide range time to digital converter," in *IEEE Nuclear Science Symposium Conference Record*, 2008, pp. 3434–3439.

[81] A. M. Amiri, M. Boukadoum, and A. Khouas, "A multihit time-to-digital converter architecture on FPGA," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 3, pp. 530–540, Mar. 2009.

[82] A. Aloisio, P. Branchini, R. Giordano, V. Izzo, and S. Loffredo, "High-precision time-to-digital converter in a FPGA device," in *IEEE Nuclear Science Symposium Conference Record*, Oct. 2009, pp. 290–294.

[83] C. Favi and E. Charbon, "A 17ps time-to-digital converter implemented in 65nm FPGA technology," in *FPGA '09: Proceeding of the ACM/SIGDA international symposium on Field programmable gate arrays*. New York, NY, USA: ACM, 2009, pp. 113–120.

[84] R. Salomon and R. Joost, "BOUNCE: A new high-resolution time-interval measurement architecture," *IEEE Embedded Systems Letters*, vol. 1, no. 2, pp. 56–59, Aug. 2009.

[85] M.-A. Daigneault and J. P. David, "Towards 5ps resolution TDC on a dynamically reconfigurable FPGA (abstract only)," in *FPGA '10: Proceedings of the 18th annual ACM/SIGDA international symposium on Field programmable gate arrays*. New York, NY, USA: ACM, 2010, pp. 283–283.

[86] C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, "A 128x128 single-photon imager with on-chip column-level 10b time-to-digital converter array capable of 97ps resolution," *IEEE International Solid-State Circuits Conference. Digest of Technical Papers.*, pp. 44–594, Feb. 2008.

[87] C. Favi and E. Charbon, "Techniques for fully integrated intra-/inter-chip optical communication," *45th ACM/IEEE Design Automation Conference*, pp. 343–344, Jun. 2008.

[88] C. Favi and E. Charbon, "A 17 ps resolution, temperature compensated time-to-digital converter in FPGA technology," Under review.

[89] R. D. Barton and M. E. King, "Two vernier time-interval digitizers," *Nucl. Instrum. Methods*, vol. 97, no. 359–70, 1971.

[90] R. G. Barton, "The vernier time-measuring technique," in *IRE*, 1957, pp. 21–30.

[91] D. R. Hoppe, "Time interpolator," *US Patent*, no. 4,439,046, 1982.

[92] D. R. Hoppe, "Differential time interpolator," *US Patent*, no. 4,433,919, 1982.

[93] M. J. Loinaz and B. A. Wooley, "A CMOS multichannel IC for pulse timing measurements with 1mV sensitivity," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 12, pp. 1339–1349, Dec. 1995.

[94] Y. Arai and T. Ohsugi, "TMC-a CMOS time to digital converter VLSI," *IEEE Transactions on Nuclear Science*, vol. 36, no. 1, pp. 528–531, Feb. 1989.

[95] M. Lampton and R. Raffanti, "A high-speed wide dynamic range time-to-digital converter," *Review of Scientific Instruments*, vol. 65, no. 11, pp. 3577–3584, 1994. [Online]. Available: http://link.aip.org/link/?RSI/65/3577/1

[96] E. Raisanen-Ruotsalainen, T. Rahkonen, and J. Kostamovaara, "An integrated time-to-digital converter with 30-ps single-shot precision," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 10, pp. 1507–1510, Oct. 2000.

[97] M. Mota and J. Christiansen, "A high-resolution time interpolator based on a delay locked loop and an RC delay line," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 10, pp. 1360–1366, Oct. 1999.

[98] T. E. Rahkonen and J. T. Kostamovaara, "The use of stabilized CMOS delay lines for the digitization of short time intervals," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 8, pp. 887–894, Aug. 1993.

[99] S. Henzler, S. Koeppe, D. Lorenz, W. Kamp, R. Kuenemund, and D. Schmitt-Landsiedel, "A local passive time interpolation concept for variation-tolerant high-resolution time-to-digital conversion," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 7, pp. 1666–1676, Jul. 2008.

[100] S. Henzler, S. Koeppe, W. Kamp, H. Mulatz, and D. Schmitt-Landsiedel, "90nm 4.7ps-resolution 0.7-LSB single-shot precision and 19pJ-per-shot local passive interpolation time-to-digital converter with on-chip characterization," in *IEEE International Solid-State Circuits Conference. Digest of Technical Papers.*, Feb. 2008, pp. 548–635.

[101] S. S. Junnarkar, M. Purschke, J.-F. Pratte, S.-J. Park, P. O'Connor, and R. Fontaine, "An FPGA-based, 12-channel TDC and digital

signal processing module for the RatCAP scanner," in *IEEE Nuclear Science Symposium Conference Record*, vol. 2, Oct. 2005, pp. 919–923.

[102] P. Sedcole, J. S. Wong, and P. Y. K. Cheung, "Modelling and compensating for clock skew variability in FPGAs," in *FPT*, 2008, pp. 217–224.

[103] L. Carrara, C. Niclass, N. Scheidegger, H. Shea, and E. Charbon, "A gamma, x-ray and high energy proton radiation-tolerant CIS for space applications," in *IEEE International Solid-State Circuits Conference. Digest of Technical Papers.*, Feb. 2009, pp. 40–41,41a.

[104] D. V. O'Connor and D. Phillips, *Time-correlated single photon counting*. London: Academic Press, 1984.

[105] W. Becker, *Advanced time-correlated single photon counting techniques*. New York: Springer, 2005.

[106] R. Rigler and E. Elson, *Fluorescence correlation spectroscopy, theory and applications*. Berlin Springer-Verlag, 2001, iSBN 3-540-67433-0.

[107] H. Qian and E. L. Elson, "Fluorescence correlation spectroscopy with high-order and dual-color correlation to probe nonequilibrium steady states," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2828–2833, 2004. [Online]. Available: http://www.pnas.org/content/101/9/2828.abstract

[108] D. L. Boiko, N. J. Gunther, N. Brauer, M. Sergio, C. Niclass, G. B. Beretta, and E. Charbon, "A quantum imager for intensity correlated photons," *New Journal of Physics*, vol. 11, no. 1, p. 013001, 2009. [Online]. Available: http://stacks.iop.org/1367-2630/11/i=1/a=013001

[109] D. L. Boiko, N. J. Gunther, N. Brauer, M. Sergio, C. Niclass, G. B. Beretta, and E. Charbon, "On the application of a monolithic array for detecting intensity-correlatedphotons emitted by different source types," *Opt. Express*, vol. 17,

no. 17, pp. 15 087–15 103, 2009. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-17-17-15087

[110] C. Niclass, M. Sergio, and E. Charbon, "A CMOS 64×48 single photon avalanche diode array with event-driven readout," in *Proceedings of the 32nd European Solid-State Circuits Conference*, Sep. 2006, pp. 556 –559.

[111] C. Niclass, "Single-photon image sensors in CMOS: Picosecond resolution for three-dimensional imaging," Ph.D. dissertation, EPF-Lausanne, Switzerland, 2008.

[112] C. Veerappan, J. Richardson, R. Walker, D. Li, M. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, "A 160x128 single-photon image sensor with on-pixel 55ps 10bit time-to-digital converter," *IEEE International Solid-State Circuits Conference, Digest of Technical Papers.*, vol. 1, Feb. 2011, to appear.

[113] C. Niclass, A. Rochas, P. Besse, and E. Charbon, "A CMOS single photon avalanche diode array for 3D imaging," *IEEE International Solid-State Circuits Conference Digest of Technical Papers.*, vol. 1, pp. 120–517, Feb. 2004.

[114] C. Niclass and E. Charbon, "A single photon detector array with 64x64 resolution and millimetric depth accuracy for 3D imaging," *IEEE International Solid-State Circuits Conference. Digest of Technical Papers.*, vol. 1, pp. 364–604, Feb. 2005.

[115] C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, "A 128 × 128 single-photon image sensor with column-level 10-bit time-to-digital converter array," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 12, pp. 2977–2989, Dec. 2008.

[116] C. Niclass, C. Favi, T. Kluter, F. Monnier, and E. Charbon, "Single-Photon Synchronous Detection," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 7, pp. 1977–1989, Jul. 2009.

[117] J. W. Goodman, F. J. Leonberger, S.-Y. Kung, and R. A. Athale, "Optical interconnections for VLSI systems," *Proceedings of the IEEE*, vol. 72, no. 7, pp. 850–866, Jul. 1984.

[118] C. Debaes, A. Bhatnagar, D. Agarwal, R. Chen, G. A. Keeler, N. C. Helman, H. Thienpont, and D. A. B. Miller, "Receiver-less optical clock injection for clock distribution networks," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 9, no. 2, pp. 400–409, Mar. 2003.

[119] A. V. Mule, E. N. Glytsis, T. K. Gaylord, and J. D. Meindl, "Electrical and optical clock distribution networks for gigascale microprocessors," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 10, no. 5, pp. 582–594, Oct. 2002.

[120] J. Fujikata, K. Nose, J. Ushida, K. Nishi, M. Kinoshita, T. Shimizu, T. Ueno, D. Okamoto, A. Gomyo, M. Mizuno, T. Tsuchizawa, T. Watanabe, K. Yamada, S. Itabashi, and K. Ohashi, "Waveguide-integrated Si nano-photodiode with surface-plasmon antenna and its application to on-chip optical clock distribution," *Applied Physics Express*, vol. 1, no. 2, p. 022001, 2008. [Online]. Available: http://apex.ipap.jp/link?APEX/1/022001/

[121] K. Ohashi, K. Nishi, T. Shimizu, M. Nakada, J. Fujikata, J. Ushida, S. Torii, K. Nose, M. Mizuno, H. Yukawa, M. Kinoshita, N. Suzuki, A. Gomyo, T. Ishi, D. Okamoto, K. Furue, T. Ueno, T. Tsuchizawa, T. Watanabe, K. Yamada, S.-i. Itabashi, and J. Akedo, "On-chip optical interconnect," *Proceedings of the IEEE*, vol. 97, no. 7, pp. 1186–1198, Jul. 2009.

[122] S. J. Walker and J. Jahns, "Optical clock distribution using integrated free-space optics," *Optics Communications*, vol. 90, no. 4–6, pp. 359–371, 1992. [Online]. Available: http://www.sciencedirect.com/science/article/B6TVF-46FR0MC-96/2/ab8ab4a68cdaed8ab1c153d3968f8195

[123] S. T. Tewksbury and L. A. Hornak, "Optical clock distribution in electronic systems," *The Journal of VLSI Signal Processing*, vol. 16, no. 2, pp. 225–246, Jun. 1997. [Online]. Available: http://www.springerlink.com/content/J856061440751173

[124] L. C. Kimerling, "Silicon microphotonics," *Applied Surface Science*, vol. 159–160, pp. 8–13, 2000. [On-

line]. Available: http://www.sciencedirect.com/science/article/B6THY-40S0D60-11/2/ac0058eebfebc33b49027e39f5a07b09

[125] P. J. Delfyett, D. H. Hartman, and S. Z. Ahmad, "Optical clock distribution using a mode-locked semiconductor laser diode system," *Journal of Lightwave Technology*, vol. 9, no. 12, pp. 1646–1649, Dec. 1991.

[126] M.-Y. Kim, D. Shin, H. Chae, and C. Kim, "A low-jitter open-loop all-digital clock generator with two-cycle lock-time," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 10, pp. 1461–1469, Oct. 2009.

[127] H. Kojima, S. Tanaka, and K. Sasaki, "Half-swing clocking scheme for 75% power saving in clocking circuitry," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 4, pp. 432–435, Apr. 1995.

[128] S. C. Chan, K. L. Shepard, and P. J. Restle, "Uniform-phase uniform-amplitude resonant-load global clock distributions," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 102–109, Jan. 2005.

[129] L. Zhang, B. Ciftcioglu, M. Huang, and H. Wu, "Injection-locked clocking: A new GHz clock distribution scheme," in *IEEE Custom Integrated Circuits Conference*, 2006, pp. 785–788.

[130] L. Zhang, B. Ciftcioglu, and H. Wu, "A 1V, 1mW, 4GHz injection-locked oscillator for high-performance clocking," in *IEEE Custom Integrated Circuits Conference*, 2007, pp. 309–312.

[131] J. A. McNeill, "Jitter in ring oscillators," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 6, pp. 870–879, Jun. 1997.

[132] M. Combes, K. Dioury, and A. Greiner, "A portable clock multiplier generator using digital CMOS standard cells," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 7, pp. 958–965, Jul. 1996.

[133] M. Z. Straayer and M. H. Perrott, "A multi-path gated ring oscillator tdc with first-order noise shaping," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1089–1098, Apr. 2009.

[134] J. Borremans, J. Ryckaert, C. Desset, M. Kuijk, P. Wambacq, and J. Craninckx, "A low-complexity, low-phase-noise, low-voltage phase-aligned ring oscillator in 90 nm digital CMOS," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 7, pp. 1942–1949, Jul. 2009.

[135] C. Favi and E. Charbon, "Techniques for fully integrated intra-/inter-chip optical communication," *45th ACM/IEEE Design Automation Conference*, pp. 343–344, Jun. 2008, supplemental material.

[136] C. Favi and E. Charbon, "Optically-clocked instruction set extensions for high efficiency embedded processors," Under review.

[137] E. G. Friedman, "Clock distribution networks in synchronous digital integrated circuits," *Proceedings of the IEEE*, vol. 89, no. 5, pp. 665–692, May 2001.

[138] S. Dasgupta and A. Yakovlev, "Comparative analysis of GALS clocking schemes," *IET Computers Digital Techniques*, vol. 1, no. 2, pp. 59–69, Mar. 2007.

[139] M. Krstic, E. Grass, F. K. Gurkaynak, and P. Vivet, "Globally asynchronous, locally synchronous circuits: Overview and outlook," *IEEE Design Test of Computers*, vol. 24, no. 5, pp. 430–441, Sep. 2007.

[140] J. Muttersbach, T. Villiger, and W. Fichtner, "Practical design of globally-asynchronous locally-synchronous systems," in *Sixth International Symposium on Advanced Research in Asynchronous Circuits and Systems*, 2000, pp. 52–59.

[141] K. Y. Yun and R. P. Donohue, "Pausible clocking: a first step toward heterogeneous systems," in *IEEE International Conference on Computer Design: VLSI in Computers and Processors*, 1996, pp. 118–123.

[142] X. Fan, M. Krstic, and E. Grass, "Analysis and optimization of pausible clocking based GALS design," in *IEEE International Conference on Computer Design*, 2009, pp. 358–365.

[143] M. Heath and I. Harris, "A deterministic globally asynchronous locally synchronous microprocessor architecture," in *4th International Workshop on Microprocessor Test and Verification: Common Challenges and Solutions*, 2003, pp. 119–124.

[144] G. Wolrich, E. McLellan, L. Harada, J. Montanaro, and R. Yodlowski, "A high performance floating point coprocessor," *IEEE Journal of Solid-State Circuits*, vol. 19, no. 5, pp. 690–696, Oct. 1984.

[145] W. Marwood and A. P. Clarke, "A coprocessor with supercomputer capabilities for personal computers," in *Proceedings of the 1988 IEEE International Conference on Computer Design: VLSI in Computers and Processors*, 1988, pp. 468–471.

[146] Y. Liu and S. Furber, "A low power embedded dataflow coprocessor," in *IEEE Computer Society Annual Symposium on VLSI*, 2005, pp. 246–247.

[147] A. Hodjat, D. Hwang, L. Batina, and I. Verbauwhede, "A hyperelliptic curve crypto coprocessor for an 8051 microcontroller," in *IEEE Workshop on Signal Processing Systems Design and Implementation*, 2005, pp. 93–98.

[148] M. D. Galanis, G. Dimitroulakos, and C. E. Goutis, "Performance improvements in microprocessor systems utilizing a coprocessor data-path," in *Embedded Computer Systems: Architectures, Modeling and Simulation, International Conference on*, 2006, pp. 85–92.

[149] N. Clark, H. Zhong, and S. Mahlke, "Processor acceleration through automated instruction set customisation," in *Proceedings of the 36th Annual International Symposium on Microarchitecture*, San Diego, Calif., Dec. 2003, pp. 129–40.

[150] L. Pozzi and P. Ienne, "Exploiting pipelining to relax register-file port constraints of instruction-set extensions," in *Proceedings of the International Conference on Compilers, Architectures, and Synthesis for Embedded Systems*, San Francisco, Calif., Sep. 2005, pp. 2–10.

[151] L. Pozzi, K. Atasu, and P. Ienne, "Exact and approximate algorithms for the extension of embedded processor instruction sets," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. CAD-25, no. 7, pp. 1209–29, Jul. 2006.

[152] P. Nagaraju, K. Anshul, and P. Kolin, "Application specific datapath extension with distributed I/O functional units," in *Proceedings of the 20th International Conference on VLSI Design*, Bangalore, India, Jan. 2007.

[153] A. K. Verma, P. Brisk, and P. Ienne, "Rethinking custom ISE identification: A new processor-agnostic method," in *Proceedings of the International Conference on Compilers, Architectures, and Synthesis for Embedded Systems*, Salzburg, Sep. 2007, pp. 125–34.

[154] A. K. Verma, P. Brisk, and P. Ienne, "Fast, quasi-optimal, and pipelined instruction-set extensions," in *Proceedings of the Asia and South Pacific Design Automation Conference*, Seoul, Korea, Jan. 2008, pp. 334–39.

[155] K. Atasu, O. Mencer, W. Luk, C. Özturan, and G. Dünda, "Fast custom instruction identification by convex subgraph enumeration," in *Proceedings of the 19th International Conference on Application-specific Systems, Architectures and Processors*, Leuven, Belgium, Jul. 2008, pp. 1–6.

[156] T. Kluter, "Architectural support for coherent architecturally visible storage in instruction set extensions," Ph.D. dissertation, EPF-Lausanne, 2010.

[157] T. Kluter, P. Brisk, P. Ienne, and E. Charbon, "Speculative DMA for architecturally visible storage in instruction set extensions," in *CODES+ISSS '08: Proceedings of the 6th IEEE/ACM/IFIP international conference on Hardware/Software codesign and system synthesis*. New York, NY, USA: ACM, 2008, pp. 243–248.

[158] T. Kluter, P. Brisk, P. Ienne, and E. Charbon, "Way stealing: cache-assisted automatic instruction set extensions," in *DAC '09: Proceedings of the 46th Annual Design Automation Conference*. New York, NY, USA: ACM, 2009, pp. 31–36.

[159] D. Lampret, "OpenRISC 1200 IP core specification," Website, 2001, http://www.opencores.org.

[160] Altera, "Nios II Custom Instruction User Guide," Website, Apr. 2010, http://www.altera.com/literature/ug/ug_nios2_custom_instruction.pdf.

[161] T. Olsson, P. Nilsson, T. Meincke, A. Hemam, and M. Torkelson, "A digitally controlled low-power clock multiplier for globally asynchronous locally synchronous designs," in *IEEE International Symposium on Circuits and Systems*, vol. 3, 2000, pp. 13–16.

[162] M. Gersbach, D. Boiko, M. Sergio, C. Niclass, C. Petersen, and E. Charbon, "Time-correlated two-photon fluorescence imaging with arrays of solid-state single photon detectors," *European Conference on Lasers and Electro-Optics, and the International Quantum Electronics Conference. CLEOE-IQEC.*, p. 1, Jun. 2007.

[163] M. Gersbach, D. L. Boiko, C. Niclass, C. C. H. Petersen, and E. Charbon, "Fast-fluorescence dynamics in nonratiometric calcium indicators," *Opt. Lett.*, vol. 34, no. 3, pp. 362–364, 2009. [Online]. Available: http://ol.osa.org/abstract.cfm?URI=ol-34-3-362

[164] D. Stoppa, F. Borghetti, J. Richardson, R. Walker, L. Grant, R. Henderson, M. Gersbach, and E. Charbon, "A 32x32-pixel array with in-pixel photon counting and arrival time measurement in the analog domain," in *Proceedings of ESSCIRC*, Sep. 2009, pp. 204–207.

[165] M. Gersbach, "Single-photon detector arrays for time-resolved fluorescence imaging," Ph.D. dissertation, EPF-Lausanne, Switzerland, 2009.

[166] E. Charbon, "CMOS single-photon systems for bioimaging applications," in *Biophotonics*, ser. Biological and Medical Physics, Biomedical Engineering, L. Pavesi and P. M. Fauchet, Eds. Springer Berlin Heidelberg, 2008, pp. 239–248. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-76782-4_13

[167] M. W. Fishburn and E. Charbon, "System tradeoffs in gamma-ray detection utilizing SPAD arrays and scintillators," *IEEE Transactions on Nuclear Science*, vol. 57, no. 5, pp. 2549–2557, Oct. 2010.

[168] T. Okusawa, Y. Sasayama, M. Yamasaki, and T. Yoshida, "Readout of a scintillating-fiber array by avalanche photodiodes," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 440, no. 2, pp. 348–354, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/B6TJM-3YF3WHY-8/2/6761ff639d4d5a6e0399ec482aa2a234

[169] K. Deiters, M. Diemoz, N. Godinovic, Q. Ingram, E. Longo, M. Montecchi, Y. Musienko, S. Nicol, B. Patel, D. Renker, S. Reucroft, R. Rusack, T. Sakhelashvili, A. Singovski, I. Soric, J. Swain, and P. Vikas, "Investigation of the avalanche photodiodes for the CMS electromagnetic calorimeter operated at high gain," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 461, no. 1-3, pp. 574–576, 2001, 8th Pisa Meeting on Advanced Detectors. [Online]. Available: http://www.sciencedirect.com/science/article/B6TJM-430WX8M-5H/2/22a4946eb0c41a6d0f5fedb86faa445e

[170] S. Vasile, R. Wilson, S. Shera, D. Shamo, and M. Squillante, "High gain avalanche photodiode arrays for DIRC applications," *IEEE Transactions on Nuclear Science*, vol. 46, no. 4, pp. 848–852, Aug. 1999.

[171] E. Charbon, L. Carrara, C. Niclass, N. Scheidegger, and H. Shea, *Radiation Effects in Semiconductors*. CRC Press, 2010, ch. 2, pp. 31–48.

[172] N. Gisin, G. Ribordy, W. Tittel, and H. Zbinden, "Quantum cryptography," *Rev. Mod. Phys.*, vol. 74, no. 1, pp. 145–195, Mar 2002.

[173] T. Jennewein, U. Achleitner, G. Weihs, H. Weinfurter, and A. Zeilinger, "A fast and compact quantum random number generator," *Review of Scientific Instruments*, vol. 71, no. 4, pp.

1675–1680, 2000. [Online]. Available: http://link.aip.org/link/?RSI/71/1675/1

[174] A. Stefanov, N. Gisin, O. Guinnard, L. Guinnard, and H. Zbinden, "Optical quantum random number generator," *Journal of Modern Optics*, vol. 47, no. 4, pp. 595–598, 2000.

[175] ID Quantique, "Quantis true random number generate," Online, Retrieved 2010. [Online]. Available: http://www.idquantique.com/true-random-number-generator/products-overview.html

[176] M. Stipčević and B. M. Rogina, "Quantum random number generator based on photonic emission in semiconductors," *Review of Scientific Instruments*, vol. 78, no. 4, p. 045104, 2007. [Online]. Available: http://link.aip.org/link/?RSI/78/045104/1

[177] J. F. Dynes, Z. L. Yuan, A. W. Sharpe, and A. J. Shields, "A high speed, postprocessing free, quantum random number generator," *Applied Physics Letters*, vol. 93, no. 3, p. 031109, 2008. [Online]. Available: http://link.aip.org/link/?APL/93/031109/1

[178] J. Richardson, L. Grant, and R. Henderson, "Low dark count single-photon avalanche diode structure compatible with standard nanometer scale CMOS technology," *IEEE Photonics Technology Letters*, vol. 21, no. 14, pp. 1020–1022, Jul. 2009.

[179] European Commision Future and Emerging Technology, "Megaframe project," Website, 2010, http://www.megaframe.eu.

About the Author

Claudio Favi was born in Milano, Italy, in 1980 where he grew up until he moved to Switzerland at the age of 8. In 1997, while still in high school, he obtained fifth position in the National Physics Olympiads. He graduated his master studies in Communication Systems in the faculty of Computer and Communication Sciences at EPFL in 2003. He spent his third bachelor academic year at Carnegie Mellon University, Pittsburg, USA as an exchange student. Claudio was also intern at the Zürich Research Lab for 6 months in 2002 where he developed new tools for IP network switch architectures validation. His master thesis was with the Center for Advanced Internet Architectures at Swinburne University of Technology, Melbourne, Australia experimenting with IP networking analysis.

After obtaining his master's degree, Claudio worked for Transports Lausannois (TL), which is Lausanne's public transportation company, as intern on the m2 project in the field of automatisms for subway transportation; while at the same time teaching object oriented programming to first year students in communication systems at EPFL.

Claudio joined Professor Charbon's group in 2005 where he pursued his PhD in the Laboratory of Processor Architectures. His research focused on the applications of single-photon sensors known as SPADs in digital CMOS technology.

On July 1st 2010, Claudio joined Nagravision SA in Cheseaux, Switzerland where he currently works as a hardware engineer on smart-card based conditional-access systems.

Claudio's interests are in hardware design, embedded system, design automation, sports, and music. He is married with two children and speaks French, Italian, and English fluently.

LIST OF PUBLICATIONS

1. **C. Favi** and E. Charbon. Techniques for Fully Integrated Intra-/Inter-Chip Optical Communication. *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*, pages 343–344, June 2008.

2. C. Niclass, **C. Favi**, T. Kluter, M. Gersbach, and E. Charbon. A 128x128 Single-Photon Imager with Column-Level 10b Time-to-Digital Converter Array. *Proceedings of the IEEE Solid-State Circuits Conference*, pages 107–13, Feb. 2008.

3. C. Niclass, **C. Favi**, T. Kluter, M. Gersbach, and E. Charbon. A 128x128 Single-Photon Imager with Column-Level 10b Time-to-Digital Converter Array. *IEEE Journal of Solid-State Circuits*, 43(12):2977–89, Jul. 2008.

4. C. Niclass, **C. Favi**, T. Kluter, F. Monnier, and E. Charbon. Single-Photon Synchronous Detection. In *Proceedings of the 34th European Solid-State Circuits Conference*, pages 114–17, Sept. 2008.

5. C. Niclass, **C. Favi**, T. Kluter, F. Monnier, and E. Charbon. Single-Photon Synchronous Detection. *IEEE Journal of Solid-State Circuits*, 44(7):1977–89, 2009.

6. **C. Favi**, R. Beuchat, X. Jimenez, and P. Ienne. From Gates to Multi-Processors Learning Systems Hands-on with FPGA4U in a Computer Science Programme. In *WESE '09: Proceedings of the 2009 Workshop on Embedded Systems Education*, pages 56–63, New York, NY, USA, 2009. ACM.

7. **C. Favi** and E. Charbon. A 17ps Time-to-Digital Converter Implemented in 65nm FPGA Technology. In *FPGA '09: Proceeding of the ACM/SIGDA international symposium on*

Field programmable gate arrays, pages 113–120, New York, NY, USA, 2009. ACM.

8. **C. Favi** and E. Charbon. A 17 ps Resolution, Temperature Compensated Time-to-Digital Converter in FPGA Technology, 2010 Under review.

9. **C. Favi** and E. Charbon. Optically-Clocked Instruction Set Extensions for High Efficiency Embedded Processors, 2010 Under review.

ACKNOWLEDGEMENTS — REMERCIEMENTS — RINGRAZIAMENTI

In this section, I'd like to address my most sincere gratitude to the many people I shared these last five years with. My life has been a mix of three languages that I've been trying to cleanly separate without success. Therefore, the following paragraphs reflect this failed trichotomy...

Il mio primo ringraziamento è per Colui che mi accompagna da sempre e mi fà ricordare che siamo solo polvere.

Sandra, Marco, e Emma, siete gli amori della mia vita. Sandra merci pour ta patience aux moments cruciaux et pour ton intolérance aux instants moins cruciaux. Je t'aime. Marco e Emma, avete illuminato con gridi di gioia e pianti tanti momenti di questi ultimi due anni. Siete tutti e due i miei tesori.

Edoardo, grazie per l'opportunità che mi ha dato. Nel 2005, quando cercai un posto di dottorante, il mio desiderio era di tendere verso un lavoro pratico tra informatica e elettronica. "Par hasard", caddei davanti a suo ufficio e lo stage di 6 mesi si trasformò in 5 anni di tesi. La tengo in piu alta stima per la sue capacità a gestire dei progetti tecnologici estremamente complessi, interessanti, con applicazioni le piu varie, e a tendere sempre verso mete ai limiti del possibile. Grazie anche per quando, nel 2006, iniziai a dubitare e che prese il tempo necessario per rimettermi in strada.

Ringrazio la mia famiglia, papà Palamede, mamma Giovanna, papà bis Camillo, nonna Elda, nonna Nina, e i miei fratelli Giordano e Jonathan. Grazie per l'amore e la stabilità offerti durante tutti questi anni di studi e per i vostri sacrifici. A ma famille par alliance, Sylvain et Agnès, Mickaël et Zoé, Maryline et Vanessa, merci de m'avoir acceuilli parmi vous. En passant de la famille par alliance à la famille spirituelle, j'aimerais souligner le soutien des pasteurs Roland O., Patrick R., Nathanaël De K. de l'église Lazare ainsi que

Marc B., Olivier G., Patrick C., Eric L., Daniel et Anita D., les personnes du groupe louange Cécilia et Oliver, Jonathan et Aurélie, Marie-Claire, Rita, Yves et Nicole, Linda et Tom, et les autres qui sont trop nombreux pour pouvoir les citer.

Many thanks to the thesis' jury, Prof. De Micheli, Prof. Wang, Prof. Jarron and Dr. Mattavelli for the comments and possible improvements before the defense and making me at ease during the defense.

The place where these 5.5 years were spent was the lab, the Laboratoire d'Architecture des Processeurs (LAP). With its staff and rolling (or sticking...) students, it has been a place of fun and hard work. Paolo, grazie di aver accettato quel gruppo Charbon di cui facevo parte e per il tuo sostegno durante questi anni. Chantal, tu as été et tu es encore le sourire du labo. Merci pour la bonne humeur que tu mets autour de toi... On doit te nommer co-directrice du labo :) Aux amateurs de café et de mots-croisés, André G., Marek, Alain, Christian, Chantal, René, Michael, et Xavier, merci pour votre accueil et les temps de détente. Merci aussi aux ingénieux Peter et André de l'ACORT ainsi que Rodolphe pour leur support technique et humain. Many thanks to the people of the Charbon's group at EPFL as well as in Delft: Claudio B., Joëlle, Christiano, Dmitri, Yuki, Matt and the others...

I'd like to address a special thanks to Theo who painfully had to bear with me the most. The 6 months of internship were under his supervision and I later moved into the INF134 office with him. I think that the many engineering skills I learned during these years where mostly from you. Thank you, I hope you got something in return besides the annoying noob questions...

Aux amis proches, Raoul et Stefanie, Nicolas et Audrey, Sylvain et Marie, Joël et Evodie, Jérôme B., Christophe et Tanja, merci pour les moments où vous ne m'avez pas posé la fatidique question: "C'est quand que tu finis ton doctorat?"... et je vous pardonne pour

toutes les autres fois. :P

Enfin, merci à mes nouveaux collègues de NagraVision, Jérôme, Alex, Roan, Cyril, Deborah, Louis, Alessandra, Marco, Raphaël, Sacha, Sylvain, Jean-Philippe, Karl et Nicolas qui m'ont accepté parmi eux alors que je finalisais la rédaction et préparais la défense de cette thèse.